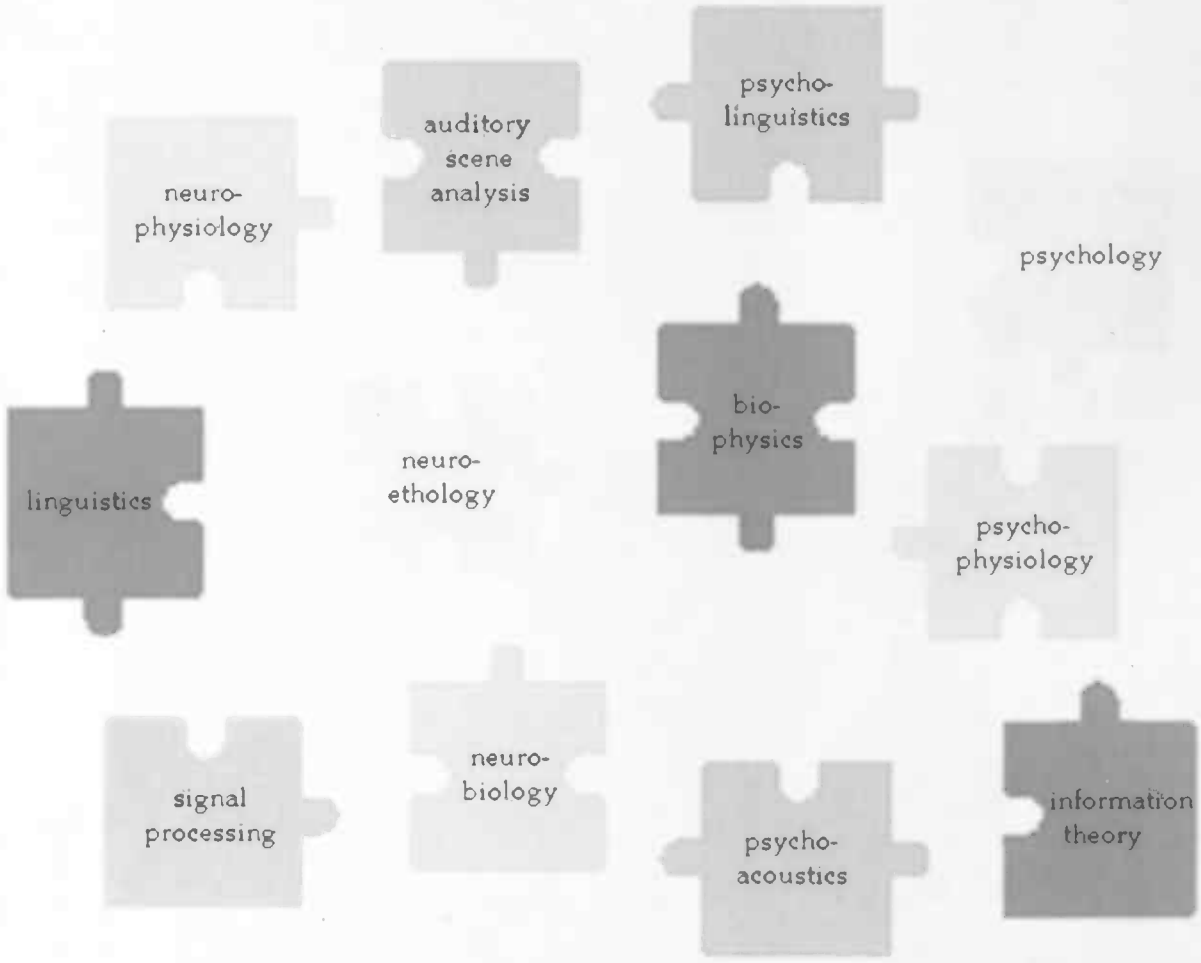
The coherent perception of speech within Cognitive Science



Desirée Houkema (0794309)

Rijks*universiteit* Groningen
Afstudeeropdracht TCW / SWI

Begeleiders:
Drs. T. Andringa
Dr. G.J. Dalenoort

Oktober 2001

The coherent perception of speech within Cognitive Science

Desirée Houkema (0794309)

Rijks*universiteit* Groningen
Afstudeeropdracht TCW / SWI

Begeleiders:
Drs. T. Andringa
Dr. G.J. Dalenoort

Oktober 2001

To Mart'n, Melvin and Ian...

Index

Chapter 3 The Auditory System 37

Chapter 4 The Auditory Scene 71

Preface

This report presents the results of a literature research project that I performed to fulfill my study requirements for both *Cognitive Science and Engineering* (TCW) and *Social-Scientific Information-Technology* (SWI), which I studied at the *University of Groningen* (RuG). It has been eight years since I started studying Psychology. After the first year, at the end of 1994, I choose to graduate in TCW and SWI. Within TCW, the idea that inspired me is the complementary interaction between Cognitive Science (CS) and Technology: knowledge obtained within CS can be used to solve or guide technical problems, and, on the other hand, engineering techniques can be used to implement and test models within CS. In my opinion, the basis on CS was minimal within TCW. This was due to its more practical approach, thereby emphasizing on the skills needed to design and implement computational systems displaying cognitive capabilities, not necessarily on understanding human cognition. Within SWI, I specialized in *Cognitive Modeling*, where the emphasis lies on modeling mental functioning based on cognitive theories. Furthermore, the application of knowledge about cognition in the field of *Human-Computer Interaction* (HCI) also formed an important area of interest. This strengthened my scientific background in CS.

During the time course of following these studies, I often felt the desire to integrate the fragmented pieces of knowledge within both studies. It therefore became an objective I wanted to achieve within my study period. Especially within the TCW courses, there was little space for such knowledge integration. The only course in which this aspect was somewhat emphasized took place in the final year, within the course *Capita Cognitiewetenschap*, in which a book on an actual issue within CS is discussed. In the year that I followed this course, a book on the dynamic nature of cognition (as opposed to a computational approach) was choosen. This formed an important contribution to my vision on cognition (see section 1.3). Fortunately, within SWI, there were several courses (given by Gert-Jan Dalenoort and/or Pieter de Vries) in which students were trained in understanding the nature of cognitive systems by looking at it from different perspectives. This was mainly achieved by making students familiar with the framework of *General Systems Theory* (GST, see also section 1.2), which is very suitable for thinking multi-disciplinary.

The final stage of SWI includes a literature research and a practical training within a company, or within the University. It was possible to combine these two with the final research project of TCW in one large project aimed at integrating the acquired knowledge in the context of a particular field of application, namely that of speech recognition. I got the opportunity to do this within the company *Human Quality (HuQ) Speech Technologies*. Within this company, new (cognitively-inspired) techniques and methodologies are developed that are suitable for application in the field of *Automatic Speech Recognition* (ASR). Since the staff of HuQ wanted to ground the cognitive basis for their existing techniques and for the development of new techniques, this company provided a perfect environment for me to finish my study and to realize the goals I set myself. I am very grateful to the HuQ staff, Tjeerd Andringa and Peter van Hengel, for their patience and flexibility, thereby allowing me to perform this research project.

This research project not only forms the completion of my study, but with it, of course, also a closure of an important period of my life. Within the last 4½ years, after 4 years of being "only" a student, I became the mother of two suns, Melvin and Ian, now 4 and 2 years old. Looking back at this period, I must say that it has not always been easy to realize the goals I set myself. The combination of taking care of Melvin and Ian, my relationship with my life partner Marten, and finishing my study requirements, would not have been possible without the enormous support I got from the people around me throughout this period. Therefore, although I can never thank them enough, I would like to take this opportunity to acknowledge their important contribution. First of all, I would like to thank my parents for always being there for me, and, in particular my mother, for taking care of Melvin and Ian when I had to study. I can think of no one else who could have done this better, and with so much love and dedication. Furthermore, I would like to thank my brother René for stepping in every now and then, and for his critical view in our frequent discussions on what is important in life. I wish more people were like him. Also, although Melvin and Ian are probably not aware of it themselves, I would like to thank them for keeping me motivated to finish my study, and at the same time reminding me, on a daily basis, where my priorities were. Last, but definitely not least, I want to thank Marten for his patience, his love and his support throughout the whole period, which was undoubtedly as intensive for him as it was for me. Thank you, Mart'n, and I promise not to do it again... (Mart'n, 'k bin wiis mei dy jer, en ik sil 't net wer dwaen...)

Desirée Houkema
Emmeloord, Oktober 2001

Chapter 1 *Introduction*

1.1 Cognitive Science and Automatic Speech Recognition

Cognitive Science (CS) is aimed at understanding the nature and functioning of the mind, both biological and artificial. It studies the biology, evolution and development of cognitive systems, investigating the mechanisms underlying cognitive abilities such as perception, recognition, information storage and information retrieval, language acquisition, comprehension and production, concept acquisition, problem solving, and reasoning. It is claimed to be an interdisciplinary field, with its own methodologies, in which knowledge from philosophy, psychology, linguistics, neuroscience, and computer science is integrated. These separate scientific fields have all approached some basic questions posed by the nature of mental processes in their own ways as part of the broader endeavor of their respective fields.

Besides the theoretical component to reach a coherent understanding of cognition, CS is also aimed at applying this knowledge such that intelligent systems capable of performing cognitive tasks can be build. This way, theoretical models, that have been formalized enough to be implemented, can be tested and further improved upon. However, from an engineering point of view, it might be argued that it is not important to understand how the system performs a certain cognitive task, as long as the system's performance satisfies the practical demands. In other words, engineers put performance first, and in a sense, are not interested in what can be learned from studying human cognitive capabilities.

When looking at the field of Automatic Speech Recognition (ASR), the approach that has been taken typically reflects such an engineering point of view. This is not surprising. Though speech recognition involves almost all of the aspects of cognition that have just been mentioned, there does not exist a "theory of speech recognition" within CS that specifies and formalizes all the relevant aspects that are involved in accomplishing this task. If the purpose is to build a working system, this is exactly what is needed. One fundamental difference between engineers and cognitive scientists is that engineers look at the whole system, whereas CS still deals with relatively isolated aspects of the system, which are often studied under unnatural, controlled conditions.

Though the description of CS given earlier suggests that the interdisciplinary approach would have led to an integrated theory of speech recognition, this ideal is still far from reality. In fact, Speech Science (as a subdiscipline within CS) is still a *multi*disciplinary field, consisting of much fragmented knowledge from many different disciplines, each discipline[1] using its own methodologies and dealing with its own problems. Given the fact that

[1] Though the Speech Sciences also consist of interdisciplinary fields such as psycholinguistics, psychoacoustics, neurolinguistics, etc., Speech Science in itself is not an interdisciplinary field as will be exemplified clearly in this overview. Even the interdisciplinarity within these subfields is in fact more an ideal than common practice. For instance, within psycholinguistics many researchers are just linguists using methodology from experimental psychology, or psychologists working with linguistic stimuli. Actually *knowledge integration* does not exist to the extent that it should, i.e., the interdisciplinarity does not exist within the participating researchers, which essentially forms the basis of a true interdisciplinary field. Furthermore, I consider the disciplines just mentioned as scientific disciplines of their own, since the integration between all these fields regarding speech recognition as a cognitive capacity should take place within CS.

many aspects are involved in the process of recognizing speech, it is impossible to study all these aspects all at once. However, in a sense, it is also impossible to study isolated aspects, or to isolate aspects, of speech recognition. This problem is neither new nor specific for the Speech Sciences, and holds for any complex system that is being studied, and therefore for many scientific disciplines. Thus, to understand the whole system it is necessary to understand its constituting parts, but to understand these parts it is necessary to understand the whole system.

So, where do we go from here? Ideally, knowledge derived from the Speech Sciences and from what engineers have discovered in building ASR systems should complement each other. However, since these disciplines have converged from one another a few decades ago, the question is whether this is really possible. The ASR approach is not aimed at building systems performing tasks in a similar way as humans perform these tasks. Furthermore, given the fact that, after decades of ASR research, the performance of current ASR systems is still far from human performance, it is obvious that engineers can still learn something from human speech recognition which is (in contrast to ASR systems) flexible, robust, and efficient (Pols, 1997; Allen, 1994, 2000).

On the other hand, Speech Science is not aimed at building working systems dealing with real-life stimuli and circumstances, and the multidisciplinary approach does not directly aid in reaching a coherent theory (or model) of speech recognition. The problems that have been encountered (and sometimes solved) by ASR engineers can therefore be used as a starting point for more realistic and more useful research questions, and for possible revisions of current speech recognition models which can then be tested (Huckvale, 1996).

Apparently, what is needed, is a reconvergence of engineers and speech scientists, such that what has been learned in both fields can be used to search for new directions in both fields (Huckvale, 1997b). This means that engineers should be more interested in how humans recognize speech, and speech scientists should be more aimed at trying to understand the whole system dealing with real-life circumstances. For ASR engineers, this might imply that, contrary to current practice, they should be prepared to temporarily accept increasing speech recognition error rates (Bourlard et al., 1996). It is not to be expected that taking a new approach will immediately lead to a level of performance that is comparable to or better than current ASR performance that has resulted from more than 20 years of optimization.

For speech scientists, it is also time to reflect on current practice. As has been mentioned earlier, knowledge about speech recognition is very much fragmented, but more importantly, it has been obtained by studying isolated aspects of speech recognition within unnatural experimental contexts, by using unspeechlike stimuli, or by using speech stimuli with nonhuman mammals. This does not necessarily mean that the obtained knowledge is of no value if we want to apply it in the broader context of understanding the human speech recognition system as a whole. However, it is important to realize that the conclusions that have been derived are, in principle, only valid in the particular context in which the research questions have been addressed. They depend on the formulated research questions, the experimental paradigm, the nature of the stimuli, the specific task demands, the theoretical framework within which they are explained, etc. Therefore, to reach a coherent understanding on speech recognition, it is necessarily that speech scientists take these specific contexts into account. A possible way to accomplish this, is by making use of the framework of General Systems Theory (GST). In the next section some ideas from GST will be shortly introduced (for more details, see Dalenoort and de Vries, 1995a).

1.2 General Systems Theory

Systems theory deals with general aspects of systems, i.e., their properties, behavior, interactions, changes, and development. In general, a system can be either open (i.e., subject to influences from outside the system) or closed,[2] but in both cases it is considered to be a coherent whole. Biological systems are always open. They are influenced by their environment, and in turn can themselves influence the environment. Additionally, they are able to adapt their internal structure and/or their internal state to the requirements of their environment.

The investigation of a certain system, starts with the discovery of the system. A system can consist of several subsystems, that individually also form a coherent whole, depending on the level of description. Thus, a system can be described at different *levels of aggregation*. There does not necessarily need to be a one-to-one correspondence between a level of aggregation, and a level of description. A multiplicity of perspectives and descriptional levels, corresponding to multiple models, to "explain" the observed system behavior[3] is permitted, as long as they are consistent with one another in a given context. To *explain* the system's behavior, it is important to have a *description of its behavior* that is as complete as possible, such that hypotheses and models can be constructed that account for the *functionality* of the internal processes leading to this behavior.

Ideally, the goal is to postulate a sufficiently formal processing model about the internal state- and process variables at a functional level to explain and *predict* the system's behavior. Finally, the goal is to specify how models or theories that explain the system's behavior at a certain level of aggregation can be related to models or theories describing aspects of the system at lower or higher levels of aggregation. Since the system behavior often emerges from the *interaction* between its subsystems, it is often neither possible nor desirable to reduce the formalized theories to one another. It is important to specify the conditions at a lower level of aggregation under which behavior at a higher level can be observed, and vice versa. These specifications provide the bridges that are needed for relating different models, and therefore knowledge from different research disciplines, to one another.

Though GST is not a theory in itself, it does provide a framework for the integration of knowledge obtained within multidisciplinary fields. Within GST, the issues that are addressed concern, for instance, the terminology that is used to describe a system, and what can and can not be expected from such a description. What does it mean to *explain* the behavior of a system, and how can explanations at different levels of description be related to one another? How has the knowledge been obtained within a certain scientific discipline? These aspects determine the validity of a certain theory or model. To integrate knowledge from different research disciplines, it is essential to critically review the experimental "facts" in the light of these and related questions.

[2] However, considering a system as closed is almost always an idealization. Within GST it is therefore emphasized what it actually means for a system to be open.

[3] Notice that the observations that describe the behavior of the system directly depend on the perspective from which the system is observed.

1.3 Dynamic systems

A specific class of systems are dynamic systems. A *dynamic system* is a system for which the status at t-1 determines the status at t. The behavior of such a system therefore evolves over time (future behavior is determined by previous behavior/history). When different dynamic systems are coupled, they can influence each other. The behavior of such a set of coupled dynamic systems as a whole can not be explained from (or reduced to) the behavior of the individual systems. What is important here, is that although the individual subsystems can be seen as *semi-independent* (as they can be defined in terms of for instance a certain difference or differential equation), when these different subsystems are coupled, the resulting behavior of this coupled system displays properties that can not be reduced to the properties of the individual subsystems. The *coupling* allows the system as a whole to display interesting, *emergent properties* that *implicitly* arise from the *interactions* between the different subsystems. It is therefore again useful to describe the larger system and its subsystems at different levels of aggregation.

Our cognitive system can also be seen as a dynamic system. For instance, it learns from previous experiences, and therefore future behavior is determined by these previous experiences. Because of this, adaptive behavior is possible. Furthermore, our cognitive system can also be considered to consist of other dynamic subsystems which again also consist of dynamic subsystems, etc. Though computationalism (based on the computer metaphor) initially played a dominating role in Cognitive Science, recent developments are also inspired by ideas from dynamic system theory which means that more emphasis is placed on the dynamics of cognition (e.g., Port and van Gelder, 1995). They therefore make use of the formalisms of dynamic system theory.[4] The behavior of our perceptual system can then be described, and ideally also explained using the terminology of dynamic system theory (e.g., attractors, bifurcations, time dependence, coupled equations, etc.). There are some interesting analogues in the observed system properties that are described within dynamic system theory, and the observed properties of cognitive systems.

For instance, cognitive processes unfold in real time. Therefore, timing always matters, and for every point in time the (mental) state of the cognitive system can be defined. The essence of dynamical models of cognition is to describe how processes unfold moment by moment, in real time. Additionally, natural cognitive systems sometimes change state in continuous ways, whereas they can also change state in ways that appear discrete. Interestingly, dynamics can deal with such discreteness in being able to describe how a continuous system can undergo changes that look discrete. Another property that also holds for cognitive systems, is that everything is simultaneously affecting everything else.

Nevertheless, given the semi-autonomous nature of the different subsystems the strength of these interactions within a system are always stronger than between systems, and not all subsystems influence one another to a similar degree. The properties that have just been mentioned allow the dynamical approach to account for the variety and interdependence of time scales at which different processes take place (for instance, by making a distinction between state variables and parameters.[5]

---

[4] It is important to note here, however, that though in dynamic system theory use is often made of difference equations (so discrete time steps are taken), it is not suggested that the underlying processes have a discontinuous nature, they are continuous (i.e., defined at every moment in time). Nevertheless, it is useful to use the formalisms (and terminology) of dynamic system theory for implementational (and descriptional) purposes, as long as the factual continuous nature of the processes being described is kept in mind and the time steps taken are small enough.

[5] State variables and parameters differ in their rate of change. Parameters can be considered as more slowly varying state variables and are therefore (relatively) constant within the time scale in which the process of interest evolves.

Another property of cognitive systems is that they are highly structured, in both their behavior and their internal spatial and temporal organization. A challenge for Cognitive Science is to describe that structure, and more interestingly, to explain how this structure could have evolved as a result of a self-organizing process. Dynamic systems are able to create structure in both space and time, and therefore organize themselves. An important critique for the validity of a cognitive model of dynamic processing has been known as the "self-organization critique", and dynamic system theory could offer a way to come up with models that can withstand this critique.

For instance, within connectionist approaches[6] to cognition, a distinction can be made between localist and distributed connectionist systems. The former places more emphasis on form and structure, whereas the latter is more inspired by how learning from previous exposure to stimuli leads to the encoding of the structure that is imposed by the system environment, in which the system's general learning capabilities is also an important determinant factor.

Though these different approaches could come up with a similar structural organization, it is also possible that a presupposed structure, which is based solely on descriptions derived from observations of a system's behavior, does not converge on the internal structure as it is represented within the cognitive system. As a result, this can pose serious problems regarding the validity of the model. Therefore, this issue of learnability or self-organization plays an important role when evaluating the validity of a certain explanatory model of cognitive processing.

This is often the case with researchers who *start* from a higher descriptional level and, subsequently, are searching for *rules* that, according to their view, should be revealed in the processing that takes place at lower levels. An illustrative example is the way linguists look at the *result* of language production, i.e., the observable product of the language production process. These observations are taken as the starting point for investigating the language processing mechanisms in the brain. As the descriptions are often rule-based (with many *ad hoc* solutions to deal with all the exceptions that can not be accounted for by such rules), they are searching for the implementation of these rules in the brain.

For example, a recently popular theory is *Optimality Theory* (OT), where the fact that the postulated rules do not always seem to hold is interpreted as being the result of the existence of weak and strong rules that can fire in the brain where the brain is searching for an optimum in the application of these rules. Though this pursuit might lead to fruitful results for simulating language behavior, there is really no a priori reason to suppose that it is possible to infer any explanatory value from observations of the product of language processing that have been obtained without referring to language as a cognitive process, and without referring to the role of language in the interaction with the environment. I believe that the search for explicit rules being represented *as such* in the brain is inconsistent with knowledge from cognitive science regarding the underlying mechanisms of information-processing and how we generally seem to structure and represent information about the environment. Furthermore, it does not take in account the role that language plays in our interaction with the environment through communication. This does not have to exclude that, when starting from a lower descriptional level, and when taking the factors mentioned into account, the behavior that can be derived (roughly) corresponds to the explicit rules that have been observed without these considerations. But, it is more likely that a two-way interaction between the different research disciplines leads to more fruitful results, because the interdependency is taken into account from the first step that starts with the formulation of the research questions.

[6] Though connectionism is compatible with dynamic systems theory, not all connectionists are dynamicists, and not all dynamicists are connectionists. Dynamic approaches within connectionism are for instance recurrent neural networks analyzed with dynamical system techniques, whereas other approaches within connectionism are sometimes more close to symbolism, and therefore more related to computationalism.

Finally, gaining a deeper understanding of the *embeddedness* of cognitive systems also plays an important role in cognitive modeling, that is, the relation of the cognitive system to its neural substrate, and the relation of the cognitive system to its essential surrounds (the body, and the physical environment). The interaction with the environment, and the ability of the system to adapt to the requirements that are imposed on it by the environment are essential if one wants to gain a deeper understanding of cognition.

1.4 Approach motivation

As has been mentioned, speech recognition involves many of the aspects of cognition mentioned in section 1.1. Understanding this process therefore requires knowledgefrom many research disciplines,[7] and the framework of GST that has been introduced in section 1.2 could be very helpful during the process of putting the pieces of knowledge together. The description of the dynamic perspective on cognition, which has been introduced in the preceding section, served to complete the general introduction, which had the purpose to clarify my general viewpoint regarding Cognitive Science.

Returning to the issue of the interaction between Cognitive Science and Engineering, more specifically regarding the issue of applying knowledge from Cognitive Science for the purpose of building ASR systems, I have chosen the following strategy. First of all, I want to take the "problem" of speech recognition as a central starting point. This means that, instead of looking directly at how certain pieces of knowledge could be applied, I first want to put these pieces together with the purpose to have a good overview of the aspects that are involved, and what is already known within CS. The purpose of this overview is to be able to give some design guidelines regarding the required system architecture of an ideal[8] ASR system. Furthermore, I want to gain insight in possibly relevant sources of information, and understand the dependencies that exist between different sources of information and how these are being processed. After the choice of the system architecture, and the nature of the representations on which the system should be able to operate, the question is which methodology can be used to implement the final system (preferably, but not necessarily, by making use of existing methodology, but this depends on the earlier choices, and not vice versa). Then, the iterative process of model formalization, system design and implementation, system testing, possible model revision, system optimization, etc. should be able to move in fruitful directions.

This strategy contrasts with the short-term approach of the majority of engineers who are building ASR systems.[9] Namely, to start with the methodology, to optimize within the limits of this methodology, and to only look for features that can be implemented on very

[7] This includes not only knowledge about the obtained results (conclusions), but also about the experimental and theoretical paradigms (methodology) within which these results have been obtained. Furthermore, to understand and appreciate certain research motivations also requires understanding the broader research context in which researchers within a certain discipline operate.

[8] The reason for using "ideal" here is that, given my background in Cognitive Science, my primary interest lies in cognitive modeling. Therefore, an ideal ASR system would be as consistent as possible with a cognitive model of speech recognition. Though I am convinced that this also leads to useful insights for the building of ASR systems, the requirements of a cognitive model and an ASR system are not the same. So, in order to be of use for ASR, some concessions related to computational efficiency, real-time performance, modeling validity versus performance accuracy, etc. necessarily need to be made. At this instant, this is not my primary concern, since I believe that this should not influence the first process of conceptualizing the nature of the problem.

[9] It should be stressed, however, that not all engineers take this approach. The idea to include perceptually-based aspects of speech recognition in ASR systems is certainly not new, and many useful advances have already been made by researchers who take this approach (e.g., Allen, 1994, 2000; Bourlard, 1996; Tchorch and Kollmeier, 1999). Unfortunately though, they usually take only one or two aspects of CS which they are subsequently trying to force into standard ASR.

short notice, such that a further (minor) decrease in error rates can be achieved as soon as possible. I believe that current ASR technology should contemplate on this strategy, since there are strong indications that there are some serious problems with the basic assumptions that have been made (see for instance, Andringa, 2001).

However, as has been argued earlier, the strategy also contrasts with approaches taken within CS, which should be more concerned with the integration of different knowledge sources in order to understand the functioning of the whole system. This also means that theories and models should be formalized in such a way that they can be implemented and tested upon real data of representative complexity. Both CS and ASR research can benefit from the knowledge that has been obtained during the past decades of research in both areas.

Therefore, from a scientific point of view, I believe in the long-term benefit of the strategy I described earlier. The purpose of this research project is therefore to make a first step in this direction by providing an overview of knowledge from CS regarding (the process of) speech recognition. Since it is neither possible nor immediately desirable to include all aspects that are involved in the process of human speech recognition, I have decided to focus on those aspects that I believe are the most promising and most useful as far as the final goal of formalizing and applying this knowledge in ASR technology is concerned. These will be concisely introduced in the next section.

1.5 Aspects of speech recognition

Speech recognition is one aspect of the larger process of speech communication in which a number of speakers of a certain language have the intention to communicate messages that contain meaningful information to one another within a certain discourse context. This intention leads to the production of a speech signal. In chapter 2 the physical aspects that are involved in this speech production process will be described, namely some of the anatomy of the human speech production system as well as how this relates to the physical properties of the speech signal. Possible signal representations that describe this physical signal will not be given in detail, since I have described this elsewhere (Houkema, 1999b). It will be assumed that the reader is familiar with signal analysis techniques such as the Fourier Transformation (see for instance, Balmer, 1991; Ifeachor and Jerwis, 1993).

Additionally, a short overview will be given of some classifications that have been made by linguists to describe the speech signal, specifically a summary will be given of knowledge from phonetics and (articulatory) phonology which is also partly based on background knowledge (e.g., O'Grady et al., 1991).

After the description of the signal properties of speech, chapter 3 will deal with what happens to this physical signal in the auditory system, after it reaches our ears, where the most important transformation takes place on the basilar membrane (BM) which is situated in the internal ear (cochlea). The neural firing pattern that results from the movements of the basilar membrane forms the initial neural code on which further processing in the auditory system is based. A description of the kind of information that can be extracted from this input signal will be given. This chapter includes knowledge from, for instance, biophysics, neurobiology, neuroethology, and neurophysiology.

Furthermore, since speech is rarely produced in isolation, one task of the auditory system is to isolate those signal components that have originated from a single source in the environment. This has been known as the problem of *Auditory Scene Analysis* (ASA) (Bregman, 1990). Humans are capable of isolating the relevant signal components in a noisy signal that result from a single speaker. This capacity makes them much more noise robust than most current ASR systems that try to match a signal representation of the whole spectrum

during a short period of time against a spectral template. In chapter 4, an illustrative overview will be given of results that have been obtained within research disciplines in which questions related to ASA have been addressed. This overview is mainly based on knowledge from psychoacoustics (a subfield of psychophysics), and psychophysiology.

Chapter 5 will deal with the speech processing system, more specifically with models of spoken word recognition, and issues that have been addressed within research disciplines such as psychoacoustics, experimental psychology, psycholinguistics, and linguistics. Here, the focus will be on how acoustic-phonetic representations are mapped onto linguistic representations such as words. It is assumed that the reader has some general background in (experimental) psychology, and is familiar with cognitive modeling techniques.

In a sense, the different chapters can be considered in relative isolation. This is because, they actually result from different subprojects within this research project. To reach the point of having an overview of the different aspects that are involved in the speech recognition process, and how they are studied and related to one another, I dealt with one issue at a time. This is reflected in the way the chapters are organized. Since I find it important to be aware of the context in which research questions are being addressed, I felt that the presentation of this overview should leave this context intact. Nevertheless, where relevant, I have added some cross-references between the different chapters. After the end of each chapter, a short summary will be given regarding the most important issues and conclusions that have been drawn throughout the chapter.

Finally, in chapter 6 a conclusion regarding the goals of this research project will be presented. Also, a reflection on the (in)consistencies between the answers that different research approaches have found while studying aspects of the speech recognition process will be given by discussing some themes that recurred throughout the different chapters.

1.6 Method

In order to write this overview, the following methods were used:

- *Background knowledge*: Chapter 2 presents an overview of background knowledge obtained within several courses regarding the physics of speech and from linguistics, and is included for completeness. For the other chapters, I also reconsulted some books that I have used for several courses.
- *Introductory books and overviews*: For chapter 3 and 4, I read some (chapters from) introductory books on hearing (e.g., Moore, 1977; Green, 1976; Allen, 1996; Zwicker and Fastl, 1999) and Auditory Scene Analysis (Bregman, 1990), in combination with (recent) articles on current research issues.
- *Literature research*: Chapter 5 is mainly based on literature research (articles), were a recent overview article from Norris et al. (2000) was taken as a starting point from which I got some idea of current issues and controverses. From this, I selected some articles that I felt were the most relevant to better understand the history of research on these issues, the used methodology and the relevance of the experimental findings. Of each article, I made a summary that included the abstract, the stimuli used, the obtained results, and the conclusions, sometimes in combination with a more detailed discussion. The different articles were kept up in a html-environment, such that it was easy to include some cross-references between different articles. While I was writing the chapter, I made use of these summaries.

- *Conferences*: I visited the following conferences:
  - The 3rd *International Conference on Cognitive Modeling* (ICCM 2000) in Groningen (23-25 March 2000). This conference is aimed at cognitive scientists who combine computational modeling with actual empirical data in their work. Usually, this is done by finding explanations of human behavior by means of computer simulations within the framework of a cognitive architecture.
  - The *Workshop "The nature of speech perception"* in Utrecht (3-7 July 2000). This was the 3rd of a series of workshops on the psychophysics of speech perception aimed at establishing a new field of interdisciplinary cooperation, called *speech psychophysics*. The goal of these workshops is to unify not only the peripheral analysis of speech and other sounds, but also the many different levels of higher-order processing involved in the perception of the speech signal, such as storage and retrieval of memory representations.
  - The 12th *International Symposium on Hearing* (ISH 2000) in Mierlo (4-9 August 2000). Here, international researchers in the areas of auditory physiology and psychology are brought together to jointly discuss current issues in hearing science.

Visiting these conferences was a very efficient way to become informed on the most recent research issues within several research disciplines, to further understand the context in which these issues are being studied, and to search for relevant literature.

Chapter 2 *The Speech Signal*

The process of speech recognition starts with a speaker having the intention to convey a message that carries information to a listener. Concisely stated, this intention is translated via a *phonological code* into a set of motor instructions which results in the production of a speech signal. In what follows some properties of this physical stimulus will be described. However, as the properties of the resulting signal are mainly determined by the physical properties of the source producing the signal (i.e., the vocal tract and articulators of the speaker), an overview of the speech production process will be given first (section 2.1).

Furthermore, an overview of linguistic units of speech representation that is based on analyses from articulatory and linguistic phonetics, and phonology, will be given in section 2.2. Within these fields, researchers are aimed at reaching formal descriptions which leads to rule-based explanations to account for the observed patterns. This nicely illustrates the relation between presupposed units of representation, and the subsequent nature of the "explanations" that have been given.

After this rather abstract description of speech sounds, some of the acoustic properties of speech sounds will be summarized in section 2.3, as well as their relation to the speech production process. It will be seen that it is difficult to map these acoustic properties on the characterization of speech as it has been obtained within linguistics. This has led to approaches that attempt to directly relate the observable acoustic consequences to the characteristics of the underlying dynamical speech producing system.

Finally, the usefulness and plausibility of such an approach will be grounded by giving some illustrative examples of the link that seems to exist between the articulatory dynamics of speech production and speech perception (section 2.4).

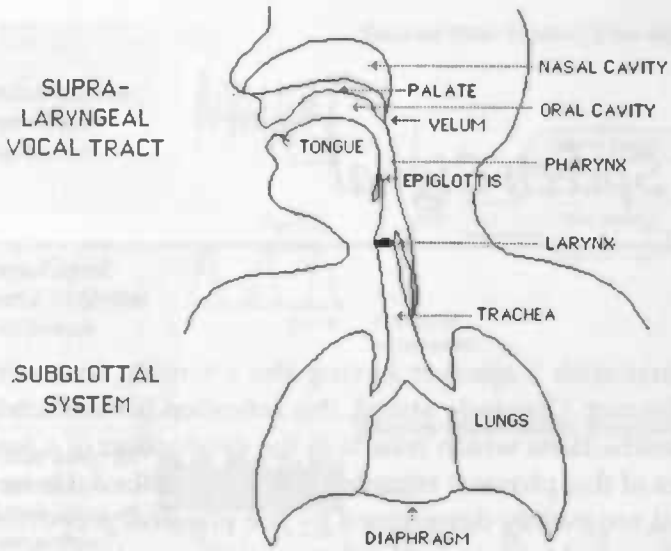


Figure 2.1 Schematic representations of the human speech production system, showing the major organs of speech.

2.1 Speech production

Speaking is in essence the by-product of a necessarily bodily process, namely the expulsion from the lungs of air after it has fulfilled its function in respiration. The organs that fulfill a role in the process of speech production primarily have a completely different function, so the role they fulfill in speech is actually a secondary. Few if any of the major organs of speech (see figure 2.1) are exclusively or even mainly concerned with speaking. The primary role of the *lungs* in respiration has already been mentioned. The role of the *larynx* (commonly known as Adam's apple) is to protect the respiratory tract, consisting of the mouth, throat and nose, against passing food. The mouth, teeth and tongue are important for taking food, and the *soft palate* and the *uvula* make sure the nasal cavity is closed while swallowing. As has been mentioned, these organs also play a role in the speak production mechanism which will be considered in what follows.

Like other acoustic stimuli the speech sound is produced when air is set in vibration, leading to variations in the air pressure waveform. The air supply in the human speech production mechanism is mainly provided by the lungs, where the *diaphragm* on the lungs produces pressure which forces air out of the lungs up the *trachea* (*windpipe*) and into the *pharynx*. The sound source that sets the air in vibration is the *larynx*.

The speech output is commonly considered to result from this source of sound energy at the larynx modulated by a *transfer function*[10] (acting as a variable acoustic filter) determined by the shape of the *supralaryngeal* vocal tract. This model is often referred to as the "*source-filter theory of speech production*" in which the combination of source and filter results in a shaped spectrum with broadband energy peaks. These are called *formants* and can be described in terms of their peak frequency, peak amplitude and bandwidth (defined by the spectral width in Hz at which the amplitude is 3 dB lower from the peak amplitude level). The shape of the filter determines the phonetic quality of the sound (see also section 2.2.2).

[10] The transfer function of a *linear* system describes what happens to a signal traveling through that system. It is often represented as a graph of gain versus frequency, which shows how much a sine wave at each frequency will be attenuated when going through the system. The peaks in the transfer function determine the formant positions, which are perceptually one of the most important features.

a) b)

Figure 2.2 View of the larynx from above showing the vocal folds and glottis. The rotation of the arytenoids cartilage pulls the vocal folds a) together (*adduction*) and b) apart (*abduction*).

Within the source-filter theory there are two sources of energy that together form the input to the filter function:

(1) a *quasiperiodic* component that results from the vibration of the *vocal folds* (also known as the *glottal source*), and
(2) an *aperiodic* (noisy) component originating at constrictions at one or more places in the vocal tract that result from the movements of the articulators (the *noise source*). A description of the production of these two sound sources will be given in what follows, and a schematic representation of the source-filter model is depicted in figure 2.3.

2.1.1 Periodic glottal source

Within the larynx is a sort of cartilaginous frame containing two important ligaments (the *arytenoid cartilage*) and muscles, which are responsible for the control of the vocal folds (*vocal cords*) to which they are attached (see figure 2.2). The vocal folds can be brought together by these structures, which is called *adduction*. This temporarily blocks the flow of air from the lungs at the *glottis* (the opening between the vocal folds) and therefore leads to increased subglottal pressure. When this pressure becomes greater than the resistance offered by the vocal folds, they are forced open again.

The pressure then drops rapidly leading to a rapid closure of the folds, as they are sucked together (also known as the Bernoulli effect). Provided that the process is maintained by a steady supply of pressurized air, the vocal folds will continue to open and close in a quasiperiodic fashion. The oscillation of the vocal folds causes puffs of airflow through the glottal opening. The frequency of these pulses determines the *fundamental frequency (F0)* of the glottal source and contributes to the perceived *pitch* of the produced sound. Actually, the signal that is produced at the glottis, is a *complex signal*, consisting also of integer multiples of the F0, called *harmonics*. The fundamental frequency depends on the mass and tension of the vocal folds and is about 110 Hz, 200 Hz, and 300 Hz for men, women and children, respectively.

In natural speech, fundamental frequency changes constantly, thereby carrying linguistic information, as in the different *intonation patterns* (pitch modulations) associated with questions and statements. It also provides information about emotional content, such as differences in speaker mood. Furthermore, the F0 pattern determines the naturalness of the produced utterance. When, for example, a synthetic version is created of a natural utterance in which the spectral properties are left largely unchanged, while the normally varying fundamental is replaced with one of constant frequency, the utterance will often not be perceived as (natural) speech.

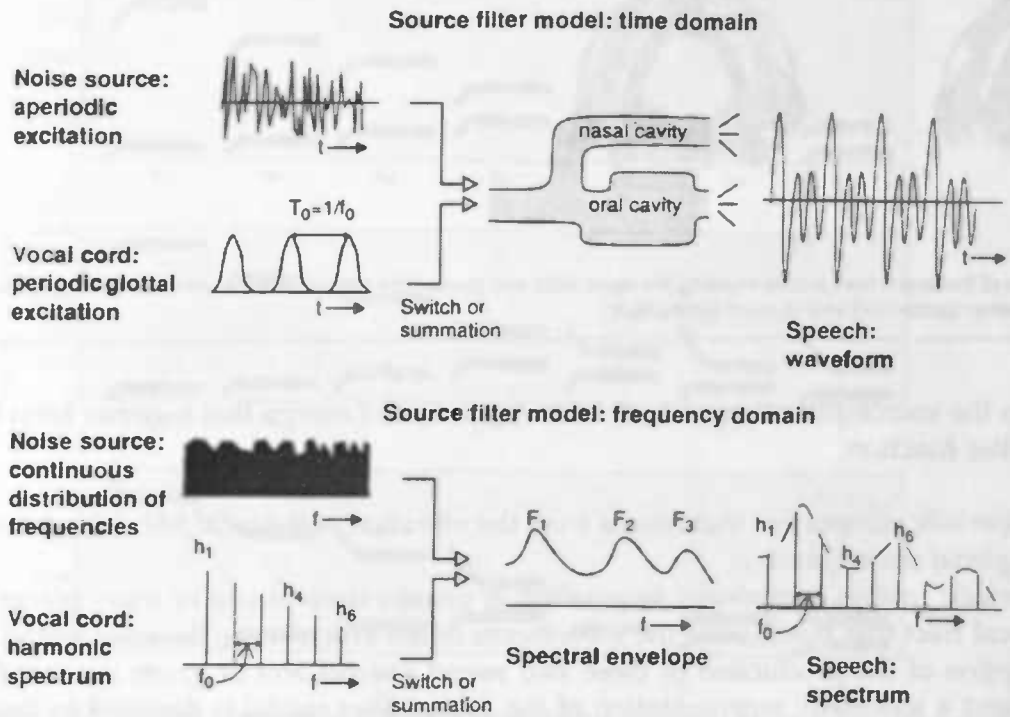


Figure 2.3 Schematic representation of the source filter model of speech production in the time domain (upper panel) and the frequency domain (lower panel). The sum of the two energy components (an aperiodic noise source and a periodic glottal source) serves as the input to the vocal tract with its filter characteristics. The formants (F1, F2, F3...) correspond to peaks in the filter transfer function, and represent the resonant frequencies of the vocal tract. The harmonics (h1, h2, ...) correspond to the frequency components that are integer multiples of the fundamental frequency (F0) of the glottal source. Notice that, in the lower panel, the amplitude of the peaks in the spectral envelope do not need to correspond to the amplitude of the highest harmonic component in a certain formant region.

At a microscopic level, there is also some *jitter* on the F0, which is important for the perceived naturalness of the speech, and for the identification of the sound as being a particular speech sound (Bregman, 1990).

2.1.2 Aperiodic noise source

The *articulators* (i.e., jaw, tongue, lips, velum, larynx) can be manipulated in various ways, thereby leading for example to the constriction of the air passage at particular places in the vocal tract. As a result of such a constriction it is possible that the sound becomes *turbulent*, which results in a turbulent or noisy component, i.e., the aperiodic noise source, in the source-filter model. Examples are the turbulence caused by constricting the tongue against the teeth (as in the "hiss" sound), or by holding the lips against the teeth (as in the /f/ in *for*). The likelihood of turbulence increases with increasing airflow, and also with decreasing constriction.

When there is a point of total momentarily closure within the vocal tract, pressure builds up behind this point leading to a sudden and abrupt release of the pressure when the closure is opened again. This causes a brief *transient* sound (e.g., the /p/ in *pay*, or the /k/ in *key*).

The manipulations mentioned thus far are mainly controlled by the jaw, tongue and lips, but the velum and larynx also have an articulatory function. The amount of air that enters the nasal passages is controlled by the *velum*, and the effective length of the vocal tract can be controlled by the *larynx height*.

2.1.3 Vocal tract filtering

Sounds where the first steps in the production process involves the oscillation of the vocal folds, are called *voiced* sounds. Sounds produced without the oscillation of the vocal folds, which can be achieved by pulling the vocal folds apart (*abduction*), are by definition *voiceless*. In either case the air from the lungs passes through the larynx into the *pharyngeal cavity* (throat), *oral cavity* (mouth) and possibly the *nasal cavity* (nose). These act as *resonators* and influence the sound in several ways, depending on the position of the articulators, which influences the size and shape of these cavities. Also, the acoustic properties of their surfaces, mainly related to the damping or energy dissipation caused by them, have an influence on the shaping of the produced sound, for these surfaces absorb acoustic energy from the air of the resonating cavities. So, the detailed shape of the transfer function is determined by the entire vocal tract serving as an acoustically resonant system combined with losses, including those due to radiation at the lips.

It is important to stress here that the pitch and formant locations can be changed independently, i.e., in the voice the coupling between source and filter is weak. By manipu-lating the voice source (at the glottis), F0 can be changed such that the voice source has the ability to oscillate at essentially any frequency. By moving the articulators, formant patterns can be changed thereby grossly modifying the final spectrum. Therefore, the sound sources and the filter can be seen as independent components in the source-filter model. This reflects a fundamental difference between the human voice and most instruments where the coupling is usually very strong.

2.2 Linguistic units of representation

The study of the sounds of human language is called *phonetics*. Phonetics deals with the configuration of the vocal tract used to produce speech sounds (*articulatory phonetics*), the acoustic properties of speech sounds (*acoustic phonetics*), and the manner of combining sounds so as to make syllables, words, and sentences (*linguistic phonetics*). The elementary speech sounds as they are usually described by phoneticians are the *phonemes*.[11] These are considered to be the basic units of sound that in any given language differentiates one word from another. Each language consists of a minimum set of phonemes, corresponding to *distinctive vocal gestures*, needed to describe every possible word in a language. This makes it possible to give a *phonetic transcription* of an arbitrary word (e.g., by means of the *International Phonetic Alphabet* (IPA)).

Not all linguists or psychologists would accept that it is appropriate to consider phonemes as "the basic units" of speech. However, phonetic spelling in terms of phonemes should not be considered as a satisfactory *acoustic* description. Phonemes are abstract units and their exact pronunciation, for example the *transition* from one phoneme to the next is left to the speaker. His good sense, along with the physical limitations on the movements of his articulators, supply this transition information.

[11] A distinction is often made between *phones* and *phonemes*. Phones are the minimal units in speech perception and production without referring to the meaning of words in a particular language in which it can occur (studied in phonetics, notation between brackets: [ ]). Phonemes are the minimal units which linguistically contrast words, these are more abstract, functional units that allow multiple phones as their realization (studied in phonology, notation between slashes: / /). Different phones that can or must function as realizations of one particular phoneme are called the *allophones* of that phoneme. As a result of different phonemic contrasts between languages, the relationship of phonemes to allophones may vary for different languages.

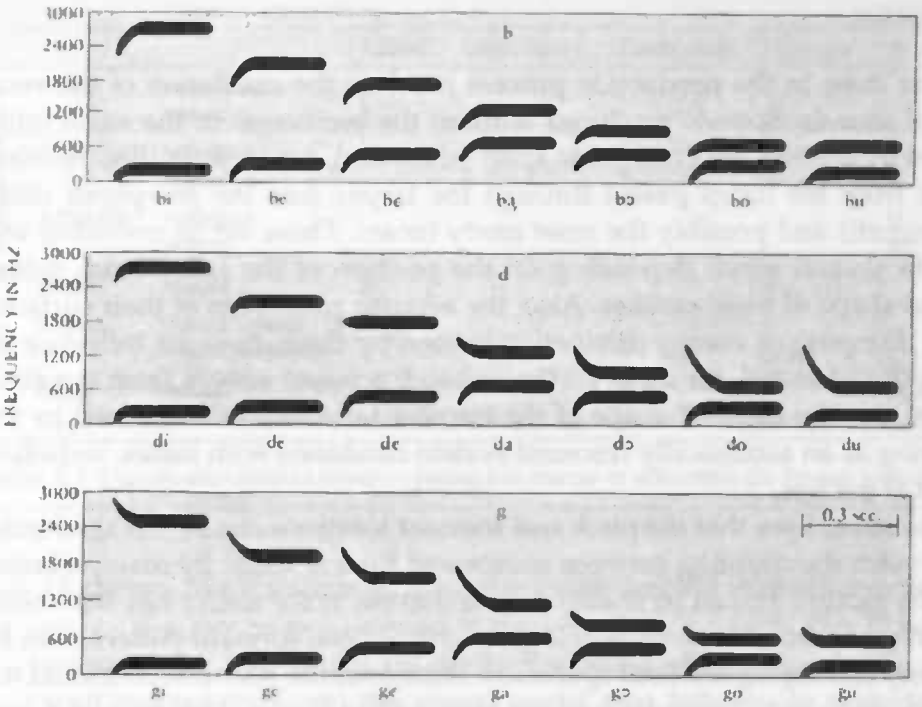


Figure 2.4 Idealized representation of syllables containing the consonants /b/, /d/ and /g/. The realization of the consonant depends on the context in which the consonant occurs, in this case the following vowel. Note that there are no invariant acoustic cues that could signal the presence of the phoneme.[12]

Phoneme transitions arise and are distinct because of *inertia* in the vocal production mechanism, so that delays introduced by the articulators are an inherent part of the speech process when we are trying to articulate a phonetic description. In the speech waveform, the transitions between phonemes seem to be as important as the elements themselves. When for example, phonemes taken from different spoken words are put together to form a new word, the resulting utterance sounds very unintelligible. Transitions also provide important cues for the perception of a speech signal as a coherent whole (see for instance section 4.9.3).

Also, a well-known effect is that of *coarticulation*. During continuous speech, the articulatory movements, in order to realize a certain target phoneme, depend on the preceding and following phonemes: the articulators are in different positions depending on the preceding one and they are preparing to the following phoneme. This causes variations on how individual phonemes are realized, called *allophones* (see figure 2.4 for some examples).[13]

The exact acoustic realization of a phoneme depends therefore (among other things) on contextual effects, speaker's characteristics, and emotions. So, it does not seem to be very useful to describe the acoustical properties (i.e., the physical realization) of speech sounds in terms of phonemes.[14]

[12] However, it has been argued that although the direction and slope of the formant transitions corresponding to the presence of the consonants are not invariant, they always point to the same frequency region for a particular consonant, which could reflect some sort of invariance, related to the resonant characteristics of the vocal tract filter.

[13] Allen (1994, 2000) states that the coarticulation effect is based upon the misconception of trying to seek for a one-to-one correspondence between a *phone* (as a physical realization of a phoneme) and a spectral template. According to him phones *can* be seen as independent units that make up higher order units such as words, as we assign evidence for the features that make up the phone in a categorically manner. This issue will be further addressed in chapter 5.

[14] The details of the phonetic transcriptions may vary however, where these different realizations of the subset of allophones belonging to a phoneme class is taken into account. Sometimes, individual symbols for different allophones are used, or use is made of diacritical signs in order to describe the specific character of a certain phoneme without having to use separate symbols.

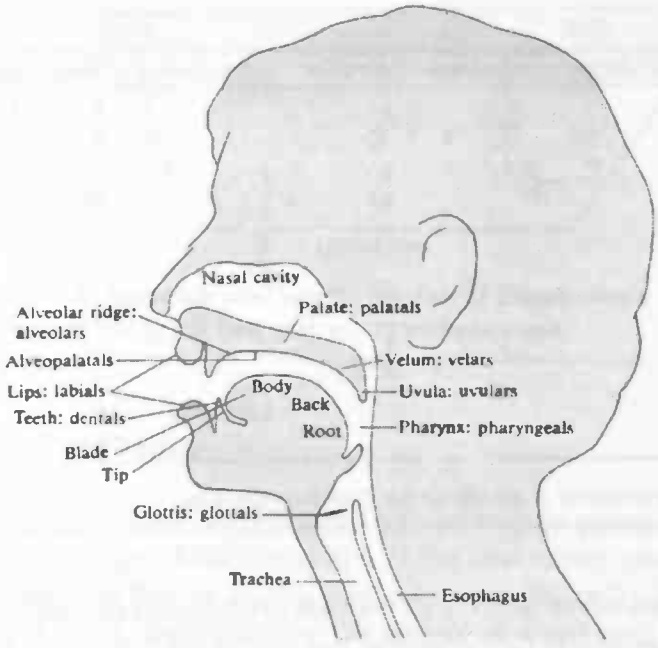


Figure 2.5 Possible places of articulation are listed, followed by the term used to describe sounds made at the place. Areas of the tongue are also designated.

Nevertheless, as the speech production system has its own built-in constraints, phonetic transcriptions are useful for the purposes for which they are devised, and the abstract notion of a phoneme is very useful for representing an important aspect of linguistic knowledge.

2.2.1 Articulatory features

Within *articulatory phonetics* a phoneme can be described according to how the sound is produced (i.e., *manner* and *place of articulation*), which is primarily studied when the phoneme is spoken in isolation. In such a way, a classificatory scheme can be developed. The main groups contained in such a scheme are the *vowels* and the *consonants*. Though there is no consensus regarding the exact definition for these two different classes, in general it can be stated that for vowels the folds are vibrating and the rest of the vocal tract is relatively open. For consonants, the vocal folds may or may not be vibrating and the tract is at one (or more) position(s) more constricted. As a result of differences in articulation, vowels and consonants differ in the way they sound. Furthermore, vowels are more *sonorous* than consonants (i.e., they are perceived as louder and longer lasting than consonants), and they have a different function in words in that they form the basis or nucleus of a syllable (which means that every syllable contains a vowel, whereas consonants are optionally present in a syllable).

2.2.1.1 Consonants

As has been mentioned, during the production of consonants the vocal folds may or may not be vibrating, so a first distinction can be made on the basis of *voicing*. Other dimensions along which the consonant class can be divided refer to aspects of the *manner* of articulation:

- *Glides*: This class of sounds shares properties of both vowels and consonants. They may be thought of as rapidly articulated vowels: they are produced with an articulation like that of a vowel.

		Place of articulation						
		Labial	Labiodental	Interdental	Alveolar	Alveopalatal	Velar	Glottal
Manner of articulation								
Stop	voiceless	p			t		k	ʔ
	voiced	b			d		g	
Fricative	voiceless		f	θ	s	ʃ		h
	voiced		v	ð	z	ʒ		
Affricate	voiceless					tʃ		
	voiced					dʒ		
Nasal	voiced	m			n		ŋ	
Liquid	voiced lateral				l			
	voiced retroflex				r, D			
Glide	voiced					j (palatal)	w	
	(voiceless)						(ʍ)	

Table 2.1 Classification scheme based on places and manner of articulation (for English consonants)

However, like consonants, they also move quickly to another articulation (as the /y/ in *yet*, or the /w/ in *wet*) or quickly terminate (for example, the /y/ or /w/ in *boy* and *now* respectively). The glides are defined in terms of their dynamic mo-vement of the vocal tract toward (or from) the vowel sound that follows (or precedes) the glide. Furthermore, they function as most consonants in the sense that they are nonsyllabic. Other terms (interchangeably) used for glides are *semivowel* or *semiconsonant*.

- *Stops*: Stop consonants (*plosives*) are produced with a complete and momentarily closure of airflow through the oral cavity which results in a transient sound. They can be voiced (/d/, /b/) or voiceless (/t/, /p/).
- *Fricatives*: These are produced with a continuous airflow through the mouth, which is a property they share with vowels and glides. They therefore belong to the larger class of sounds called *continuants*. They are characterized by a *turbulent noise* (i.e., noise associated with turbulent airflow), and may consist of that noise alone (/s/, /f/, and /x/ as in Dutch *gaan* (*go*)), or together with a periodic glottal source (/z/, /v/, and /ɣ/ as in Spanish *agua* (*water*)).[15]
- *Affricates*: These are *noncontinuants* characterized by a slow release of closure (compared to stops) from the point of articulation. Since they are characterized by turbulence together with a momentary obstruction of the flow of air through the vocal tract, they can be seen as a combination of stops and fricatives (e.g., /dʒ/ as in *judge*).
- *Liquids*: Liquids are continuants where the vocal tract constriction is not as prominent as it is for the fricative consonants. An example is the lateral /l/, where air escapes through the mouth along the lowered sides of the tongue, or the English *retroflex* /r/ (as in *ride, car*).
- *Nasals*: When the velum is lowered to allow air to pass also through the nasal cavity instead of exclusively trough the oral cavity, nasal sounds are being produced (/n/, /m/, or /ŋ/ as in *sing*).

The consonants can be further specified according to their *place* of articulation (i.e., *bilabial*, (*labio-/ inter-*) *dental*, *alveolar*, (*alveo*)*palatal*, *velar*, *uvular*, *pharyngeal*, *glottal*, see figure 2.5). A possible classification for English consonants based on the above mentioned distinctions in place and manner of articulation is shown in table 2.1.

[15] In this case the periodicity is not characterized by the presence of a set of harmonics in the speech signal, but rather by a *pattern of amplitude modulation* (AM). This means that the amplitude of frequency components is modulated periodically, see also chapter 3 and 4.

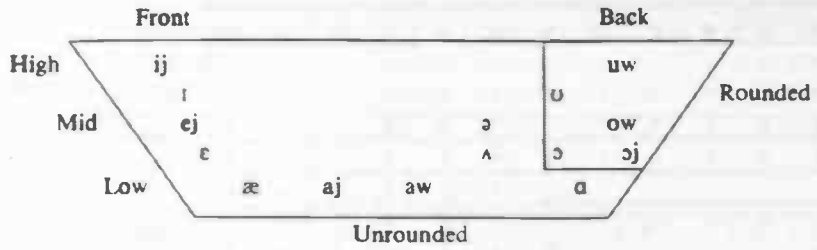


Figure 2.6 Articulation chart with the inventory of English vowels, including diphthongs, according to the position of the tongue. N.B. The [ij], [ej], [uw], [ow], and [ɑ] are tense vowels.

2.2.1.2 *Vowels*

Different vowels are produced by varying the placement of the body of the tongue and shaping of the lips. Also, the shape of the oral cavity can be further altered, for example by protruding the lips to produce *rounded* vowels, or by lowering the velum to produce a nasal vowel. And, depending on the degree of vocal tract constriction during articulation, vowels may be *tense* (produced with a relatively greater degree of constriction of the tongue body or tongue root) or *lax* (roughly the same tongue position, but with a less constricted articulation). Vowels are therefore often described with reference to tongue position (*high, low* and *back, front*), tension (*tense* or *lax*), and lip rounding (*rounded* or *unrounded*).

In (English) vowels, a further distinction can be made between simple vowels, which do not show a noticeable change in quality (e.g., *pit, put, cat*), and *diphthongs* which exhibit a change in quality within a single syllable, due to for instance tongue movement away from the initial vowel articulation toward a glide position. Diphthongs are therefore transcribed as a vowel-glide sequence (e.g. [ij] in *heat*, [aj] in *buy*). An example of an *articulation chart* with the inventory of English vowels where these differences are illustrated is shown in figure 2.6.

2.2.1.3 *Suprasegmentals*

All sounds have certain inherent *suprasegmental* or *prosodic* properties on the basis of which they can be further characterized no matter what their place or manner of articulation. These properties are *pitch, loudness*, and *length*.

- *Pitch*: All sounds give a subjective impression of being relatively higher or lower in pitch. Pitch is especially noticeable in sonorous sounds like vowels, glides, liquids, and nasals. But, even stop and fricative consonants convey different pitches. For instance, for the fricatives [s] and [ʃ], the [s] is clearly higher pitched. Pitch can be controlled by varying the tension of the vocal folds and by varying the amount of air that passes through the glottis. There are two kinds of pitch movement: *tone* and *intonation*. Within so-called tone languages, differences in word meaning are signaled by differences in pitch. Intonation refers to pitch movement that is not related to differences in word meaning, and typically takes place over a longer time span, e.g., over the course of speaking a whole utterance.
- *Loudness*: All sounds have some degree of intrinsic loudness, but this property on itself does not have a distinctive character.
- *Length*: Length is a phenomenon that is present in many languages, and refers to the fact that the articulation of some vowels or consonants are held longer relative to that of other vowels and consonants. This can also be related to differences in word meaning.

		i	ɪ	e	ɛ	æ	ʌ	ə	ɑ	u	ʊ	o	ɔ
Major class features	[consonantal]	-	-	-	-	-	-	-	-	-	-	-	-
	[sonorant]	+	+	+	+	+	+	+	+	+	+	+	+
	[vocalic]	+	+	+	+	+	+	+	+	+	+	+	+
Laryngeal feature	[voice]	+	+	+	+	+	+	+	+	+	+	+	+
Place feature	[round]	-	-	-	-	-	-	-	-	+	+	+	+
Dorsal features	[high]	+	+	-	-	-	-	-	-	+	+	-	-
	[back]	-	-	-	-	-	+	+	+	+	+	+	+
	[low]	-	-	-	-	+	-	-	+	-	-	-	-
	[tense]	+	-	+	-	-	-	-	+	+	-	+	-
	[reduced]	-	-	-	-	-	-	+	-	-	-	-	-
Manner feature	[continuant]	+	+	+	+	+	+	+	+	+	+	+	+

Table 2.2 Feature matrix for English vowels.

- *Stress*: Stress is a cover term for the combined effects of pitch, loudness, and length. Vowels that are perceived as more prominent than others are called stressed. The effect of these prosodic features varies within and between languages. An example in English is that stressed vowels are generally higher in pitch, longer, and louder than unstressed ones, but this is not always the case. The prominence typically takes place with respect to other syllables, and is usually accomplished by a relatively large shift in one or all of the three parameters pitch, loudness, and length.

2.2.1.4 *Linguistic-phonetic features*

It is also possible to characterize a phoneme by means of a *binary distinctive feature set*, which is partly related to the earlier mentioned articulatory features. Here, each phoneme is characterized by either the absence (indicated by -) or presence (indicated by +) of a certain feature. For instance, a feature matrix corresponding to the vowel [ɑ] would be:

+vocalic -consonant +sonorant	These features define the segment as consonant, vowel, or glide (here, a vowel)
-high +low +back	These features define the placement of the tongue (here, a low back vowel)
-round	This feature defines lip rounding (here, unrounded)
+tense	This feature defines tensiness/laxness (here, tense)

Examples of other feature matrices for English vowels and consonants are depicted in table 2.2. and 2.3 respectively.

Such a reference to features as the units of phonological structure that make up segments has a number of advantages. First of all, they directly reflect the underlying coordinate activities such as voicing, tongue position, lip rounding, and so on. Each feature is considered to be rooted in an *independently* controllable aspect of speech production.

Secondly, each feature may represent a phonologically relevant characteristic of segments. In other words, features express *natural classes*: classes of sounds that share a feature or features, such as voiceless stops, glides, high vowels, nasal consonants, etc. Any natural class requires fewer features to define it than to define one of its members. Therefore, it is the feature, in this case [voice], that is contrastive, i.e., the contrast between each pair of phonetic segments with otherwise identical articulations (/p/-/b/ and /k/-/g/) resides in the feature [voice], which is therefore a (language-specific!) distinctive feature.

		p	pʰ	b	t	tʰ	d	k	kʰ	g	f	v	s	z	θ	ð	ʃ	ʒ	ʧ	ʤ	m	n	ŋ	l	r	j	w	ʍ	h	ʔ
Major class features	[consonantal]	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
	[sonorant]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-
	[vocalic]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Laryngeal features	[voice]	-	-	+	-	-	+	-	-	+	-	+	-	+	-	+	-	+	-	+	+	+	+	+	+	+	+	-	-	-
	[CG]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
	[SG]	-	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
Place features	[labial]	+	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-
	[round]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-
	[coronal]	-	-	-	+	+	+	-	-	-	-	-	+	+	+	+	+	+	+	+	-	+	-	+	+	-	-	-	-	-
	[anterior]	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	-	-	-	-	+	+	-	+	+	-	-	-	-	-
	[strident]	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-
Dorsal features	[high]	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+	+	-	-
	[back]	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-
Manner features	[nasal]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-
	[continuant]	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	+	+
	[lateral]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
	[delayed release]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-

Table 2.3 Feature matrix for English consonants

Likewise, the contrast between /t/-/s/, /p/-/f/, /b/-/v/, and /d/-/z/ can be expressed with the feature [continuant].

However, these features are more than another way to represent phonetic descriptions, since they can also refer to phonological patterning. For example, the feature [coronal] refers to the class of sounds made with the tongue tip or blade raised, which is a feature that is required to state the phonological constraint on the selection of consonant sequences in coda position (the consonants to the right of the nucleus (the obligatory vocalic member) of a syllable) in English. When a vowel is tense and followed by two consonants (as in *pint*), or when a vowel is lax and followed by three consonants (as in *next*), the final consonant must always be [+coronal].

A final advantage regarding the reference to features is that it allows for the understanding of the nature of allophonic variations more exactly. In this sense, allophonic variation is not just the substitution of one sound for another, but rather the environmentally conditioned change or specification of a feature or set of features, thereby reflecting *phonotactic constraints*. Phonotactics refers to the set of constraints on how sequences of segments pattern in a language, and forms part of a speaker's phonological knowledge.

A distinction can be made between universal and language-specific phonotactics. The former refers to segment sequences that are generally considered to be unpronounceable thereby reflecting constraints on human linguistic ability. The latter refers to the fact that each language has its own set of restrictions on the phonological shapes of its syllable constituents. For example, in English, voiceless liquids and glides occur after voiceless stops in words such as *please*, *proud*, and *pure*. This results from *articulatory processes* such as *progressive assimilation*,[16] in this case *liquid-glide devoicing*: the value of the feature [voice] changes to [-voice] after voiceless consonants.

[16] *Articulatory processes* are articulatory adjustments that occur during the production of speech. Examples are assimilation, dissimilation, deletion, epenthesis, and metathesis (see for instance, O'Grady et al., 1991). *Assimilation* refers to the influence of one segment on another, and can be either *regressive* (anticipating, moving backwards to a preceding segment), or *progressive* (moving forwards to a subsequent segment). Another type of assimilation is flapping, which results in voicing and sonority being maintained throughout a sequence of segments. Though there are only a finite number of articulatory processes operating within a certain language, they produce a great deal of linguistic variability.

Within linguistics it is assumed that the unpredictable features of a certain phonemic segment are basic or underlying, such that the phonetic representation can be derived by the use of *phonological rules*. Thus, phonetic forms are derived by setting up the underlying representation and then allowing the rule or rules in question to operate in those contexts were they are relevant. The rule regarding liquid-glide devoicing can then be depicted in the following form:

$$\begin{bmatrix} \text{-vocalic} \\ \text{+sonorant} \\ \text{-nasal} \\ \text{+voice} \end{bmatrix} \rightarrow [\text{-voice}]/\sigma \begin{bmatrix} \text{-vocalic} \\ \text{+consonantal} \\ \text{-sonorant} \\ \text{-continuant} \\ \text{-voice} \end{bmatrix} \text{———}$$

This rule is read as follows: Liquids and glides become voiceless after syllable-initial voiceless stops.

To summarize, linguists view the phonetic segments as composed of smaller elements, namely linguistic-phonetic features. These binary features have a distinctive character. Since in normal speech, phonemes are never produced in isolation, phonemes differ in how they are pronounced as a result of articulatory processes. The resulting allophones can therefore change on a certain feature dimension dependent on the context in which they occur. The underlying phonetic segments can be derived via phonological rules describing predictable patterns of variation related to phonotactic constraints.

2.2.1.5 *Syllables*

A final level of phonological representation that will be discussed here is the syllable. The internal structure of a syllable can be described by four subsyllabic units. The *nucleus* is the only obligatory member of the syllable, which is always a sonorant (usually a vowel). The *coda* consists of those segments that follow the nucleus within the same syllable, which together make up the *rhyme*. Finally, the *onset* is made up of those segments that precede the rhyme in the same syllable. The reason for assuming subsyllabic units is that syllables comply with certain constraints that prohibit certain forms of syllabification within a language.

Within a given language, the following rules can be applied to set up syllables. First, the syllabic nucleus is constructed, since this forms the only obligatory syllable constituent. Further, onsets are constructed before codas, which means that the longest sequence of consonants to the left of each nucleus that is phonotactically legal makes up the syllable onset. The remaining consonants to the right of each nucleus form the coda. Since the phonotactic constraints are language-specific, this procedure can yield different syllabify-cations in different languages.[17]

A reason for treating syllables as units of phonological structure is that they are relevant to stating generalizations about the distribution of allophonic features. For instance, in English, each voiceless stop has an aspirated and an unaspirated allophone. However, when referring to syllabic structure, the feature [aspiration] can be stated generally, since the environments where aspirations occur are only syllable-initially. Another example is the phonological predictability of phonetic length in English vowels that are shorter before voiceless consonants, before sonorant consonants, and in word-final positions, and are longer before

[17] Generally though, the syllabification seems to conform sonority-based principles where segments are ranked on a hierarchy of sonority. By making use of such sonority hierarchies the most common patterns of segment organization found across languages can be captured. An example is for instance the *Sonority Sequencing Principle* (SSP) which states that more sonorous segments stand closer to a syllable peak (or nucleus) than less sonorous ones. Also, the *Minimum Distance Sonority Principle* proposes that segments combine on the basis of their relative distance on the sonority scale. However, these sonority-based principles are not without exceptions.

voiced non-sonorant consonants. This can again be generalized, namely that they are only long when followed by a voiced obstruent in the same syllable.

The above motivation for the existence of syllabic units relates to the advantage that can be obtained in structuring the results of linguistic analyses. It does not relate to speech production or speech perception. However, other reasons exist to ground the usefulness of syllables as units of representations (see for instance chapter 5, section 5.3.4). At this point, it is obvious that allophonic variations of phonemes make the classification of phonemes on the basis of phoneme-specific features not very tractable. Therefore, the fact that allophonic distributions of linguistic-phonetic features can be generalized within syllables holds a strong case for the unity of syllables that can be distinguished on the basis of phonetically-related features.

Nevertheless, the allophonic variations do not have any contrastive function within a certain language (since they do not relate to differences in meaning). Therefore, though a phoneme is never pronounced, only its allophones, phonemes still represent useful abstract linguistic knowledge. Another argument for the reality of phonemes is that it is does not seem to be a coincidence that spelling systems usually ignore non-distinctive phonetic variations by using only one letter for the allophones of a certain phoneme. Finally, it has been shown that such allophonic variations are easily ignored in perception, even though they are systematically produced in the appropriate context. For instance, the difference between the voiced and voiceless allophones of /l/ is hardly heard by English speakers.)n the other hand, phonologically relevant distinctions, such as that between /l/ and /r/ (in English), are never missed. But, in languages where these two sounds are not contrastive (such as in Japanese), /l/ and /r/ are allophones of the same phoneme, and Japanese listeners have much difficulty in making the distinction between /l/ and /r/.

2.2.2 Acoustic properties

It has been demonstrated that different phonemes can be distinguished based on differences in articulation leading to a (language-specific) *unique set of features* for *each* phoneme. An idea that motivated much research is that differences in articulation also show themselves in the physical realization of the signal, i.e., in the air vibrations that comprise and transmit sound. If these features can indeed be identified in the pressure waveform, this would allow the unique classification of a particular phoneme. The underlying idea is the assumption that our perception of phonemes is based on information provided by *invariant acoustic cues* associated with that phoneme, which has been known as the *theory of acoustic invariance*. The earlier discussion concerning the variation in the realization of a particular phoneme already indicated that this idea is too simplistic. Nevertheless, it is useful to illustrate some of the acoustic consequences that are related to the way the sound is produced.

Before doing this, however, some of the signal representations that are often used to visualize speech signals will be mentioned. First of all, a signal can be plotted in the *time domain*. This means that the amplitude development of the pressure wave is plotted as a function of time. By applying a *Short Term Fourier Transformation* (STFT)[18] the signal can also be represented in the frequency domain, so that the contribution of different frequency components within a certain time frame can be shown by plotting energy (in dB) against frequency. This is called an *energy spectrum*. Notice that phase information is not present in such a representation. These two representations were already illustrated in figure 2.1 in the context of the source-filter model.

[18] See any book on signal analysis, for instance Balmer (1991), Ifeachor & Jerwvis (1991).

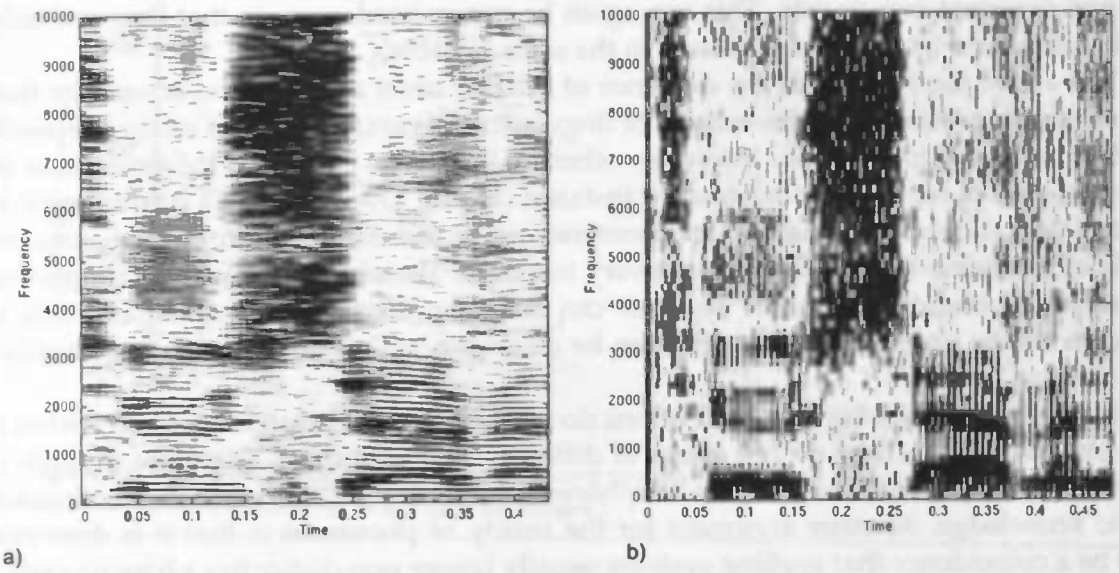


Figure 2.7 Spectrogram of the speech signal *"two seven"*. The time and frequency resolution depends on the frame size. a) With a large frame size, the transition between different sounds becomes blurred, but the fundamental frequency and harmonics are clearly separated. Therefore, the frequency resolution is quite good. b) When the frame size is smaller, the frequencies are more spread, which leads to a poorer frequency resolution. However, the time resolution is improved such that information concerning the timing is more clearly visible.

The STFT can also be used to give a time-frequency representation called a *spectrogram* where succeeding frames in time correspond to the time dimension at the x-axis, which is plotted against frequency at the y-axis. The relative contribution of each frequency component can be represented by using a color scheme for different intensities. Since the spectrogram represents both time and frequency information, this representation is used most often. However, there is a trade-off between the time and frequency resolution that depends on the frame size, i.e., the length of the time frame within which the STFT is applied. A longer frame size leads to a better frequency resolution, whereas a shorter frame size leads to a better time resolution. This is illustrated in figure 2.7.

2.2.2.1 *Voicing*

One of the most important distinguishing features between phonemes is whether or not they are voiced. Remember that when the vocal folds were oscillating, the frequency of the glottal pulses determined the fundamental frequency (F0), which in turn corresponds to the perceived *pitch*. Since the signal that is produced at the glottis, is a complex signal containing also the harmonics corresponding to the F0, this results in a harmonic (or line) spectrum when the signal is plotted in the frequency domain (see figure 2.8). The source intensity for different harmonics usually decreases by approximately 6 dB per octave.

2.2.2.2 *Vowel formants*

All vowels are produced by a glottal source that is modified through the resonant properties of the vocal tract shape that determines which frequency regions become more or less attenuated, and consequently, where peaks and valleys in the *spectral envelope* will occur. The peaks in the transfer function correspond to formant frequencies. The intensity of the individual harmonics are determined by *both* the source energy and this filter function (for examples, see figure 2.8).

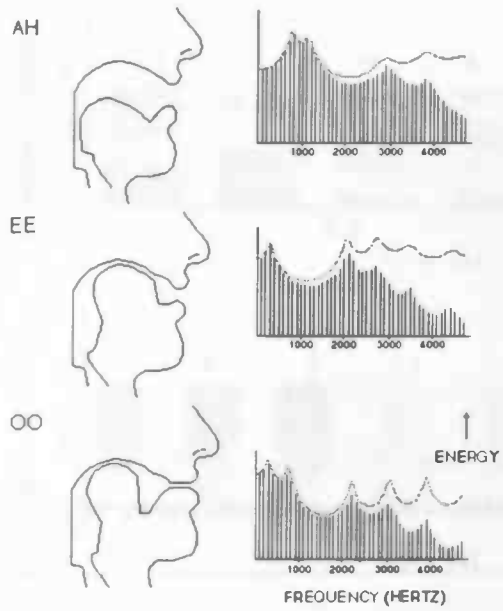


Figure 2.8 The shape of the vocal tract during the production of the vowels AH, EE , and OO (as in respectively *but*, *beet*, and *boot*), the corresponding transfer function, and the resulting energy spectrum.

For vowels, the combination of the locus of the first two formant frequencies (and the amount of energy at each of these formants) can uniquely determine the phoneme identity. Though less accurately, they can also be specified according to the tongue position. The position of the first formant (F1) corresponds roughly to the front-to-back dimension, and the second formant (F2) corresponds roughly to the height of the tongue on the low-to-high dimension. The higher the vowel, the lower F1. The more front the vowel, the higher F2. So, for example, the high front vowel /i/ has a high F2 and a low F1, while the low back vowel /a/ has a high F1 and a low F2. A plot of F1 against F2 for different vowels is given in figure 2.9. This figure shows that the vowels all fall within a triangle with the vowels /a/, /i/ and /u/ at the corners, the *acoustical vowel triangle*.

2.2.2.3 *Voiced consonants*

When looking at a speech spectrogram, it can be seen that there is little energy in the parts corresponding to the presence of consonants compared to the vowels. Furthermore, vowels have relatively stable spectra, at least for a short period of time. Some consonants, like for example fricatives, can be characterized on the basis of their formant positions, since they have relatively stable spectra just as with vowels (figure 2.10). Other consonants can be characterized by rapid changes in the frequency of a given formant or set of formants, corresponding to changes in the position of the articulators. These formant transitions are important cues for their perception. Though the occurrence of these transitions is independent of the presence of voicing, voicing does provide an important distinction between phonemes with the same places of articulation that only differ on the voicing dimension (see also section 2.2.1.4). Voicing leads to the presence of quick and regular vertical pulses in the spectrum (i.e., the sound spectrogram),[19] reflecting a pattern of amplitude modulation that is related to the fundamental frequency of the glottal pulses (related to the perceived pitch).

[19] The visibility of these glottal pulses in the spectrogram again depends on the resolution (i.e., frame size).

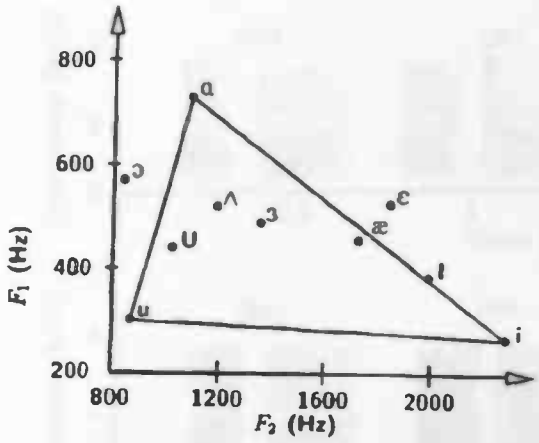


Figure 2.9 Plot of the first formant F1 versus the second formant F2 of some of the English vowels, showing the *acoustical vowel triangle* with the vowels */a/*, */i/* and */u/* at the corners .

However, the detection of voicing for consonants may not always be easy, since voiced stops and fricatives often show shorter duration times and weaker intensities than their voiceless counterparts. Furthermore, there is no harmonic structure present such that individual harmonics can be identified, and used as a basis for pitch detection.

2.2.2.4 *Consonant formants*

Voiceless fricatives (like /f/, /s/ and /x/) differ mainly in the energy distribution along the frequency spectrum. For instance, since /f/ is produced at the lips, there is an absence of a resonant cavity between the place of constriction and the point where air leaves the mouth. As a result, the energy in the spectrum is relatively equally distributed, i.e., it shows no profound peaks in the spectrum at particular frequency regions. Overall, the amount of energy is weaker compared to the /s/ and the /x/. The /s/ is produced with a relatively small, and the /x/ with a greater resonance cavity. The strongest frequency components for the /s/ therefore lie in the high frequency region, whereas for the /x/ the low frequency components are more prominent. For their voiced counterparts, the /v/ and /z/, the presence of the glottal source leads to a strong reduction of the amount of air flowing through the vocal tract. Therefore, the (aperiodic) noise that is produced at the point of constriction is much weaker and of shorter duration.

For stop consonants the aperiodic part in the frequency spectrum lies in the same frequency regions as with the fricatives just described. This is because they are produced at similar places in the mouth: /p/-/f/, /s/-/t/ and /x/-/k/ are pairs where the filtering has a comparable effect. The main difference between fricatives and stops is that for the latter the aperiodic part is relatively short and accompanied by a preceding period of silence (for voiceless stops like /p/, /t/ and /k/), or a period in which a weak periodic sound in a very low frequency region is present (for the voiced stops like /b/, /d/ and /g/). This period of (almost) silence is the result of the momentarily total closure and is shorter for voiced stops compared to voiceless stops. The pairs /p/-/b/, /t/-/d/ and /k/-/g/ have again a similar place of articulation, but they differ on the voicing dimension. Also, after the acoustic burst resulting from the release of closure, an interval of aspiration[20] is present for voiceless stops, but this interval is missing for voiced stops (see figure 2.11a).

[20] After the release of voiceless stops, a brief delay before the voicing of a following vowel can be heard. Since the lag in the onset of voicing is accompanied by the release of air, this phenomenon is called aspiration.

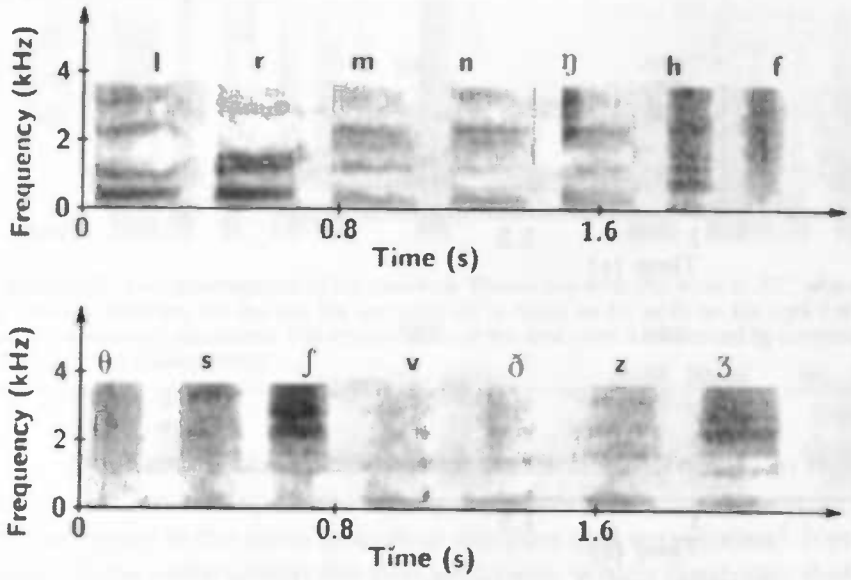


Figure 2.10 Some examples of English (relatively) steady-state consonants /l, r, m, n, ŋ, h, f, θ, s, ʃ, v, ð, z, ʒ/. For some consonants, the formant positions are visisble in the spectrogram, others are mainly characterized by their energy distribution along the frequency spectrum.

When a plosive is followed by a vowel, the length of time between the release of closure and the presence of voicing in this vowel, the *voice onset time* (VOT), is an important perceptual cue.[21] A very short VOT leads nearly always to the perception of a preceding voiced stop, whereas a very long VOT leads to the perception of a voiceless stop.

Nasals like /m/, /n/ and /ŋ/ are all voiced. The filter now consists of two parts: the oral cavity behind the point of constriction and the nasal cavity. Because there are two resonant cavities, multiple sound paths are present which in this case leads to destructive interference (*cancellations*, called *zeroes*, or *anti-resonances*) for some of the high frequency components. Nasals like /m/ and /n/ therefore show only strong frequency components in the low frequency region. In the production of the /ŋ/, the oral path is effectively cut off, because the velum is lowered and the back of the tongue is raised so far that they join. Because of this, the transfer function is only determined by the nasal cavity and there are no zeroes.

A lot of research has been aimed at finding acoustically invariant cues for speech sounds corresponding to the articulatory features described earlier. Given the relative nature of the acoustic properties corresponding to these features, it seems that they can only be determined and all have to be interpreted in the *context* of other sounds, and other characteristics of the signal, the speaker and the language. Therefore, when only a part of a spectrum corresponding to a particular phoneme is available, it is not possible to uniquely identify that phoneme. There is no one-to-one correspondence between the acoustic proper-ties of the allophonic realization of the phoneme and the spectrum at the time the phoneme is pronounced. This in itself does not pose any problem if it were possible to determine a many-to-one correspondence representing the allophones of a certain phoneme, but this is not possible, since the same sound can be perceived as a different phoneme depending on contextual factors.

[21] Possibly because it corresponds to the detection of the feature [aspiration] (see section 2.2.1).

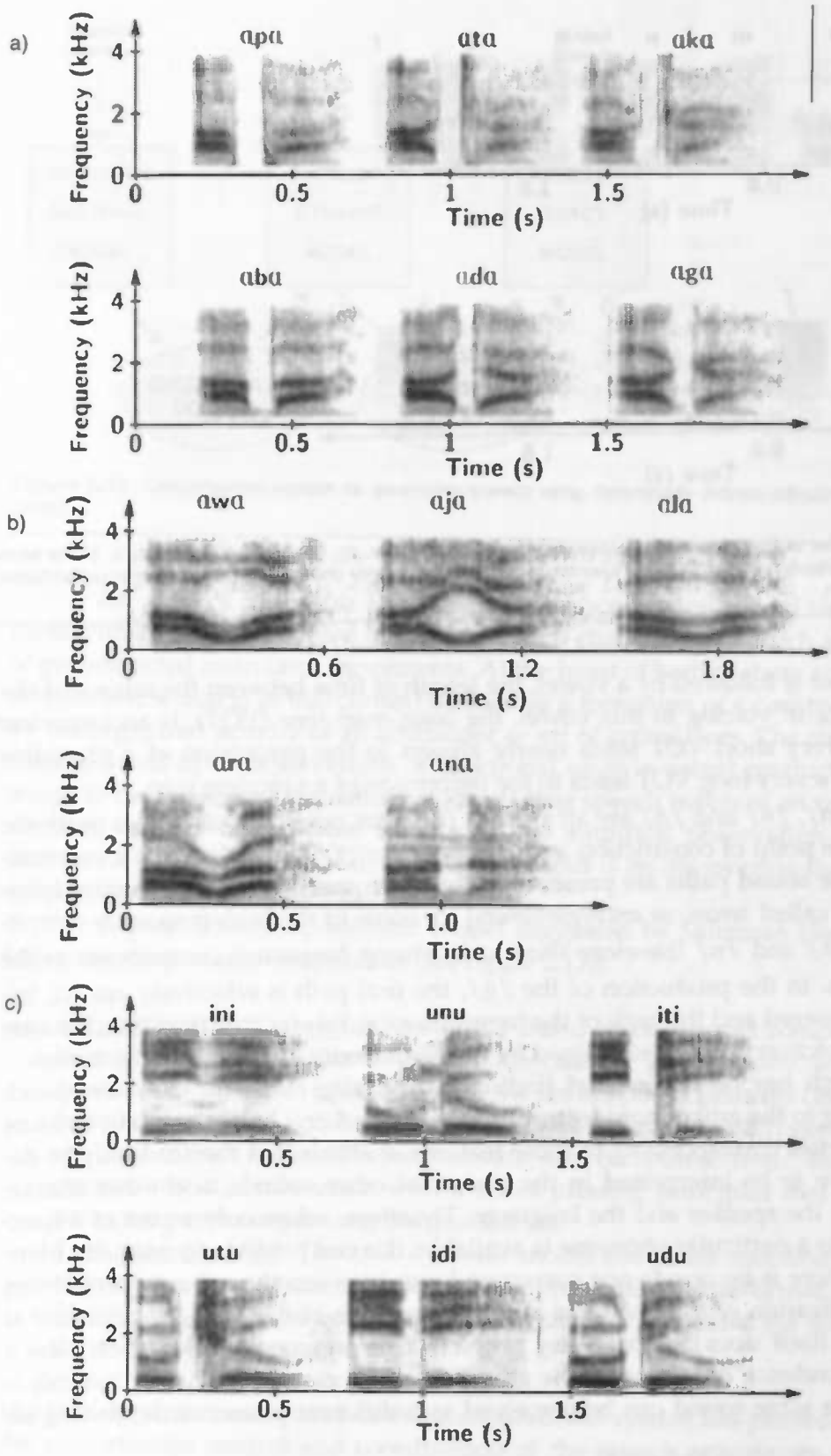


Figure 2.11 (a) Spectrograms of English stop consonants in vowel context. The upper part shows some voiceless stops /*apa*, *ata*, *aka*/, where a period of silence is visible. The lower part shows their voiced counterparts /*aba*, *ada*, *aga*/, where the period of silence is absent, and a weak periodic signal is present in the low-frequency region. (b) Spectrogram of English sonorant consonants in vocalic context: /*awa*, *aja*, *ala*, *ara*, *ana*/. (c) Spectrogram of English stops and nasals in vowel context: /ini, unu, iti, utu, idi, udu/.

Figure 2.12 Two spectrograms of the sentence *"Please say what this word is: XX"*, where the introductory sentence is spoken by different speakers. On the left, the last word ***XX*** is heard as *bit*, while on the right it is heard as *bet*. In both cases the last words are identical waveforms. The interpretation of the final word is influenced by contextual factors related to speaker characteristics. (From Slaney, 1995).

An example of this contextual variation is illustrated in figure 2.12. Two spectrograms are shown for two speakers saying the same introductory sentence, where the last word in the sentences is the same (identical samples and waveforms). Nevertheless, this word is perceived differently within the two sentences, which illustrates that recent experience changes our perceptual interpretation of words.

However, there is another problem that indicates the inadequateness of the description of phonemes in terms of independent distinctive articulatory gestures, besides the obvious influence of coarticulation effects, and other contextual effects. This pertains to the observation that most of the different articulators are coupled, in the sense that they reflect a coupled structure where the different articulators together try to reach a certain phoneme target. For instance, through *compensatory articulation* it is possible to produce a particular phoneme with different places of articulation. Within the field of articulatory phonology these aspects of phonological patterning in speech production are taken into account.

2.3 Articulatory phonology

In human communication, the speech system is specialized for the rapid transfer of information (Mattingly & Liberman, 1988). Significant events in the acoustic signal may occur in an overlapping or parallel fashion due to the coproduction of speech gestures. A result of this is that aspects of the signal corresponding to different linguistic units, such as consonants and vowels, often cannot be isolated in the acoustic stream. One way to help tease apart the components of the speech signal is to consider the physical system that gives rise to the acoustic information: the acoustic encoding of phonetic information is then viewed in light of the flexibility inherent in the production apparatus, particularly the human supralaryngeal vocal tract, in which individual articulators or *groups* of articulators can function *semi-independently*.

In the description of the speech production process it became clear that the configuration of the vocal tract, depending on the speech articulators, "shapes" speech acoustics. In trying to find a correspondence between the acoustic properties and the vocal tract configuration for the unique identification of a particular phoneme, it was obvious that there is considerable *acoustic diversity*. This has led to the desire of *articulatory simplicity*. The acoustics are continually changing during speech, so it is the behavior of the speech articulators over time, i.e., the changes of articulatory configuration and their acoustic consequences, that must be analyzed. This leads to the necessity and desirability of examining the *underlying dynamical system*.

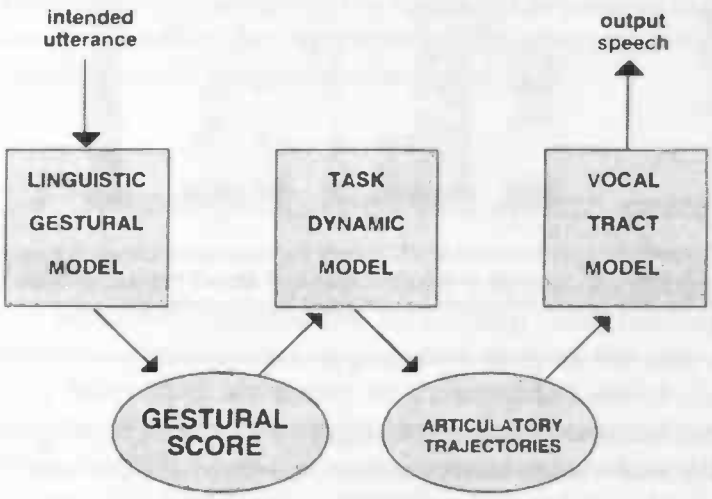


Figure 2.13 Computational system for generating speech using dynamically defined articulatory gestures. From Saltzman (1995).

An attempt to do this is by merging a phonological model based on gestural structures with an approach called *task dynamics* that characterizes speech as coordinated patterns of goal-directed articulator movements. At the heart of both of these approaches is the notion of a *gesture*, which is in this context defined as a formation of a constriction in the vocal tract by the organized activity of an articulator or set of articulators. The choice of gestural primitives is based upon observations of functional units in actual production. These models attempt to reconcile the linguistic hypothesis that speech involves an underlying sequence of abstract, context-independent units, with the empirical observation of context-dependent interleaving of articulatory movements. The focus is on discovering the regularities of gestural patterning and how they can be specified.

A particular computational model suggested by Saltzman (Saltzman, 1995) has the following three major components (see figure 2.13):

- *Linguistic-gestural model*: A gesturally-based phonological component (the linguistic-gestural model) provides, for a given utterance, a "*gestural score*" which consists of specifications for dynamic parameters for the set of speech gestures corresponding to the input phonetic string and a temporal activation interval for each gesture, indicating its onset and offset times. These intervals are computed from the gesture's dynamic parameters in combination with a set of phasing principles that serves to specify the temporal patterning among the gestural set.
- *Task dynamic model*: The task dynamic model computes coordinated articulator movements from the gestural score in terms appropriate for a particular vocal tract model.
- *Vocal tract model*: The vocal tract model, in turn, computes the speech waveform from these articulatory movements.

The task dynamic model used in this computational system has proved useful for describing the sensorimotor control and coordination of the speech articulators, as well as of skilled activities of the limbs.

For given gestures, the goal is specified in terms of independent task dimensions, called *tract variables*. Each tract variable is associated with the specific set of articulators whose movements determine the value of that variable. For example, one such tract variable is *Lip Aperture* (LA), corresponding to the vertical distance between the two lips. Three articulators can contribute to changing LA: the jaw, the upper lip, and the lower lip.

Tract variable		Articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

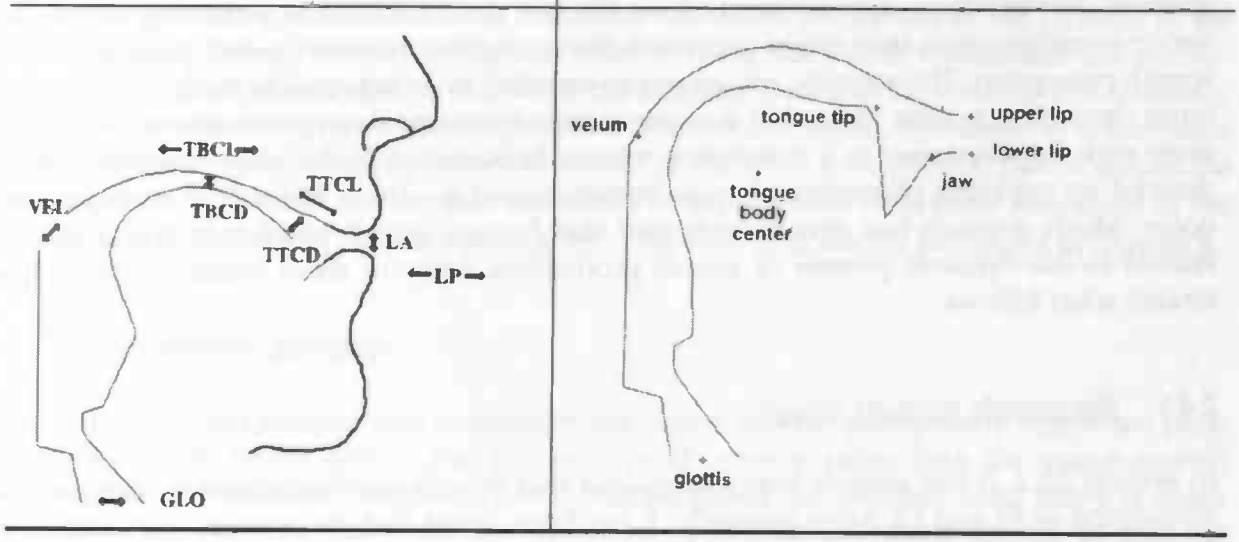


Figure 2.14 Tract variables and their associated articulators.

The standard set of tract variables in the computational model, and their associated articulators, is depicted in figure 2.14. Tract variables and articulators compose two sets of coordinates for gestural control in the model.

In addition, each gesture is associated with its own activation coordinate, whose value reflects the strength with which the associated gesture "attempts" to shape vocal tract movements at any given point in time. Invariant gestural units are posited in the form of context-independent sets of dynamical parameters (e.g., lip protrusion target, stiffness, and damping coefficients), and are associated with corresponding subsets of all three coordinate systems.

Thus, the tract-variable and model articulator coordinates of each unit specify, respectively, the particular vocal tract constriction (e.g., bilabial) and the articulatory synergy that is affected directly by the associated unit's activation. Currently the model offers an intrinsically dynamic account of interarticulator coordination within the time span of single and temporally overlapping (coproduced) gestures, under normal conditions as well as in response to mechanical perturbations delivered to the articulators, e.g., as a result of constraining the jaw.

At the present stage of development, the task-dynamic model does not provide a dynamic account of intergestural timing patterns even for simple speech sequences. Current simulations rely on explicit gestural scores to provide the timing patterns for gestural activation intervals in simulated utterances. While such explicitness facilitates research by enabling to model and test the current hypothesis of linguistically significant gestural coordination, an approach in which temporally ordered activation patterns are derived as implicit consequences of an intrinsic *serial dynamics* would provide an important step in modeling processes of intergestural relative timing. Computational modeling of connec-tionist dynamical

systems has investigated the control of sequences (e.g., Grossberg, 1986). Such serial dynamics is well-suited for orchestrating the temporal activation patterns of gestural units in a dynamical model of speech production.

2.4 Speech production and speech perception

At this point, there does not exist a satisfactory theory or model of speech production that can accurately describe the underlying vocal tract dynamics as well as the acoustic consequences. Nevertheless, approaches such as the one described in the preceding section are very interesting, since they might provide a direct coupling between speech production and speech perception. The acoustic consequences can then be interpreted in terms of the underlying dynamical system instead of a segment-based acoustic description related to discrete units such as phonemes, or a description where the so-called underlying phonemes can be derived by applying phonological rules in the context in which allophonic variations can occur. Much research has already indicated that human speech perception is intrinsically related to the dynamic process of speech production. Some of these findings will be illustrated what follows.

2.4.1 Reasonable formant ranges:

In section 2.2.2, it has already been mentioned that vowels can be defined in terms of the disposition of F1 and F2. More generally, it has been found that the position, and pattern of change, of the first three formants are the major determinants for speech perception. Analyses of the location of these formants indicate that they are confined to the lower 3-4 kHz of the spectrum which is also the range to which humans are most sensitive. If a sound has more than three formants in this frequency region, humans typically cease to perceive the sound as being speechlike. For speech perception it is important that the sound could plausibly be produced by a normal human speaker, and the formants should therefore lie within reasonable formant ranges which is determined by the physical and physiological limits of the articulators.

This is also reflected in how two vowels that are mixed together are perceived. For instance, in the production of "yah", the articulators move slowly from the /i/ to the /a/ shape, occupying the spaces between to form the intermediate vowel sounds. A likely vowel trajectory would be

/i/ - /I/ - /E/ - /æ/ - /U/ - /a/

which is obtained by making interpolations in the articulatory domain. The spectrum of the intermediate sound /E/ is quite different than a spectral interpolation of /a/ and /i/. Therefore, the sound that results when mixing the sounds together is not perceived as speech from one human speaker, since it contains six formants in the region below 3500 Hz. Nevertheless, it is possible to perceptually separate the mixed signal such that the identity of the original sounds /a/ and /i/ are recognized. As will be discussed in chapter 4, this is possible by performing a sort of auditory scene analysis that is based on phonetically-related organizational processes.

The independence between the pitch and formant locations mentioned earlier (section 2.1.3) also forms an important determinant in the interpretation of speech. For instance, when the sound of a human speaker is speeded up, the pitch and the entire spectrum also shift upwards, which is interpreted as a speech signal being produced by a person that is physically smaller than normal. This seems to be the most plausible interpretation for the fact that the formants have all shifted upward along with the pitch. The opposite effect occurs when slowing down the speech sound, which leads to the perception of speech being produced by a speaker whose head is larger than normal.

Another issue is that, in vowel perception, humans are very sensitive to the locations of the first three formants. Though the shape of the resonant formant affects the perceived naturalness of the sound, it is not as important for vowel perception as the location of the formant peak. In fact, vowel identity can even be recognized with vowels that are synthesized by using only three sine waves, one at the location of each of the three speech formants. Since the formants are not physically present in the acoustic signal, this seems to reflect some abstract sensitivity to the independent influence of the vocal tract filter on how it shapes the final spectrum. This can also explain our ability to separate the individual phonemes in the mixed signal of /a/ and /i/ described earlier. This issue will also be addressed in chapter 4.

2.4.2 Articulatory gestures

Much research has already been inspired by the notion that the perceptually significant parts of speech might be the *gestures* that produce speech sounds, rather than the speech sounds themselves. For instance, Liberman (1967, 1982) proposed the so-called *motor theory of speech perception* in which he views the perception of speech as modularly special in that the mechanisms in the brain for acoustic speech recognition are so intrinsically tied to the production mechanisms that they are essentially inseparable. According to Liberman, the motor programs built for making speech sounds and the inherent limitations of the articulators cause humans to process speech sounds through an elaborate network that resonates with familiar production patterns.

Others interpret the strong connection as reflecting innate sensitivities of the auditory system for speech motor targets (Skoyles, 1998) without assuming one underlying neural substrate. In this sense, phones serve as the means by which speech transmits the articulation information used to map speech into pronunciation. Phones therefore reflect a replication code that enables humans to engage in vocal imitation in the process of language acquisition. According to Skoyles, speech imitation requires the imitation of these motor goals. Therefore, the organization of vocal movements is interpreted in terms of motor targets, i.e., not the sequentiality of individual movements is important (since speech can articulate several movements at the same time), but the sequentiality of targets (consistent with the approach taken by Saltzman described in section 2.3, though based on different motivations). These targets refer to complex vocal tract goals such as resonance properties[22] and are considered to be perceptually discrete (i.e., either being perceived by one target or another). Again, what is important is that these imitable pronunciations must be producable by vocal tracts of different sizes and shapes thereby ruling out "template type" acoustic invariants, but instead referring to higher-order characteristics, such as rates and shape of modulation, and rates and shape of frequency shifts.

[22] This reflects very complex and abstract knowledge such as, for instance, that some speech sounds are the same because they share second formant transitions which point to a common frequency locus (examples of this type of invariance can be seen in figure 2.4).

2.4.3 Articulatory interpolation

One reason for assuming that speech understanding involves more than simple instant-to-instant acoustical analysis comes from experiments with ambiguous speech sounds, such as the /b/, /d/ and /g/. These are easily confused, since they differ most only in one formant trajectory. For example, in a /ba/, /da/, /ga/ syllable context, the essential difference between /da/ and /ga/ is in the region above 2 kHz where the third formant (F3) sweeps upward in /ga/, but downward in /da/. The F3 of /ba/ lies somewhere in between. Due to masking phenomena (see chapter 4, section 4.5), the ear can not track the F3 trajectory after each voiced plosive accurately enough to unambiguously determine the presence of either of these consonants. Because many speech sounds are ambiguous, and their final interpretation depends on many contextual factors, this has been taken to indicate that speech perception relies on cues other than the simple static acoustics of a certain speech phone.

Another interesting phenomenon comes from experiments where subjects are presented simultaneously with auditory and visual stimuli, for instance an auditory stimulus from a person saying /ga/ combined with a visual stimulus of a person making lip movements corresponding to /va/. Listeners become aware of the percept /va/ which again seems to indicate that perception takes place in terms of articulatory movements. Another example is the combination of an auditory /bi/ with a visual /gi/ leading to the intermediate percept of /di/. This can be explained by referring to the location of closure within the vocal tract: /bi/ is formed with the closure at the lip end, /gi/ with a closure back on the tongue, and /di/ is formed with a closure that lies between those points of articulation. This class of phenomena has been known as the McGurk effect (discovered by McGurk, 1968).

An advantage for viewing speech perception in the light of the underlying dynamics, is that whereas nearly all descriptions of the speech sounds exhibit discontinuities, an *articulatory descriptions is always continuous* since it directly pertains to the fact that the articulators are constrained to move smoothly from one point to the next by occupying all states in between. Our sensitivity to this underlying continuous process of speech production can be illustrated by digitally inserting silences in recorded speech signals. For instance, when a silence is inserted in the nonword *sish* between the /s/ and the /i/, and between the /i/ and the /ʃ/, most English speakers interpret the resulting sound /s_I_ʃ/ as *stitch*. They therefore favor the known word that is most likely to be uttered, assuming the continuity of the articulators. When presented with the intact sound *sish*, many listeners perceive *fish*, which is understandable given the small difference between the sounds /s/ and /f/, and the fact that *fish* is a word whereas *sish* is not. Nevertheless, no one perceives *ftitch* in the edited utterance since this does not conform to phonotactic constraints in English. This pertains to another aspect of speech perception that is based on linguistic filtering.

Interestingly, when multiple lexical interpretations are possible, as is for instance the case in *say* edited to become /s_aɪ/, the final interpretation seems to be related to articulatory simplicity. For short durations of the inserted silence, most people perceive *stay*, whereas for longer silences people perceive *spay*. This is because it takes longer to move the mouth from the shape required to form an /s/ to that required to form a /p/ (involving two different points of constriction) compared to that required to form a /t/ (involving the same point of constriction on the tongue), again indicating the presence of an articulatory component in the perception of sounds. Such a form of perceptual filtering in terms of more easily pronouncable sequences also occurs in the interpretation of nonwords where no lexical significance to the words is involved. For instance, inserting short silence in the utterances *ara* and *ala* after the consonants /r/ and /l/ causes them to be perceived most commonly as *arga* and *alda*, since the places of articulation for /r/ and /g/, and /l/ and /d/ are most similar.

2.5 Summary

Within the source-filter theory of speech production there are two sound sources: (1) the glottal source, originating from the vibration of the glottis, and resulting in a quasiperiodic component, and (2) the noise source, originating at places of (total) constriction in the vocal tract, and resulting in an aperiodic component. A particular speech sound can contain either one of these components, or both. These sound sources form the input to the filter function, whose characteristics are determined by the shape of the vocal tract. This filter shapes the final spectrum by amplifying certain frequency regions (corresponding to formants) more than others. The absence or presence of voicing, and the (relative) location and development of the formants are important characteristics of a speech signal. Vowels and some consonants can be identified by the disposition of formant peak locations (particularly F1 and F2), and other consonants are characterized by formant transitions. The formant transition is, at every instant in time, providing information about two phonemes, the consonant and the vowel – that is, the phonemes are transmitted in parallel.

A lot of research has been focused on the independent and invariant aspects of elementary speech units in the acoustical signal that roughly correspond to phonemes or phoneme features. It has been demonstrated that descriptions from articulatory phonetics that are based on the articulation of individual phonemes are inadequate for this purpose. This is because the production of speech does not consist of a discrete sequence of individual phoneme productions, but rather reflects a continuously evolving process where (a set of) articulators are trying to reach a certain motor target in the context of preceding and subsequent phonemes This leads to coarticulation effects that are intrinsically related to the physical constraints of the human speech apparatus.

Within linguistics, use is made of linguistic-phonetic features. These are language-specific, contrastive features. The reality of phonemes is based on an abstract notion of linguistic knowledge. The observation that individual phonemes are never produced, but only their allophones, has led to the formulation of phonological rules that describe how and in which context linguistic-phonetic features may change. These rules are needed to derive the assumed underlying phonemes. By referring to the syllable as a linguistic unit of representation, the phonological rules that are needed to account for the observed variability in the realization of phonemes can be formalized much more efficiently.

However, it has been argued that in order to understand the exact nature of the acoustic consequences that result from the underlying vocal tract dynamics in speech production, it is useful to describe the production, acoustics, and perception of speech with relation to the dynamics of the underlying articulatory gestures. The examples given in section 2.4 indicate that the acoustics of speech are indeed often interpreted and perceived in such a way.

Chapter 3 *The Auditory System*

In order to understand how the physical speech signal leads to the perception of the words and the meaning conveyed with it, it is important to gain a deeper understanding of the signal processing that takes place after it reaches the ear. In the *peripheral auditory system* the signal undergoes several transformations which determine the first neural representation that is available of the acoustical signal. This serves as the neural code on which the *central auditory system* (CAS) must base further processing. A perspective on the neurobiological system that allows us to couple to our acoustic environment provides insight on issues like:

- Which signal properties can be adequately represented by the auditory system?
- Where in the brain are they represented?
- How are the connections between different brain structures that represent information regarding different signal properties?
- What is the direction of information flow between these brain structures?, etc.

In what follows an overview of the neuroanatomy (structure) of the auditory system (AS) will be given. The auditory pathway from the ear to the *auditory cortex* (AC) will be discussed, though considerably more is known about the structure and function of the peripheral portions of the AS (described in section 3.1) compared to the physiology (functional role) of more central structures. Auditory research is often conducted on nonhuman mammals. Since the peripheral portions of the AS of most mammals are quite similar, these structures are probably of no specific importance for the understanding of speech processing in humans. However, transformations of the acoustic environment produced by the peripheral structures are central to the speech processing functions of the cortex.

From figure 3.1 the complex structure of the CAS can be seen. It consists of numerous cell groups (*nuclei*) within the *brainstem* with a high degree of connectivity between the nuclei. There are ascending pathways (*afferent, feedforward*) as well as descending pathways (*efferent, feedback*). The descending pathways eventually (indirectly) terminate in the periphery (see section 3.8). Some of the diverse connections undoubtedly serve a reflex function, some interact at a low level with the afferent system, and some serve as points of binaural interaction with homological fibers from the contralateral ear (Green, 1977).

Though it has long been thought that the principal function of the auditory system's brainstem nuclei is sound localization, recent studies have identified brainstem mechanisms which possibly play an important role in the perception of speech. These structures are responsible for multiple representations of the stimulus which are conveyed to the cortex in multiple parallel functional pathways, representing spectral and temporal properties of the signal.

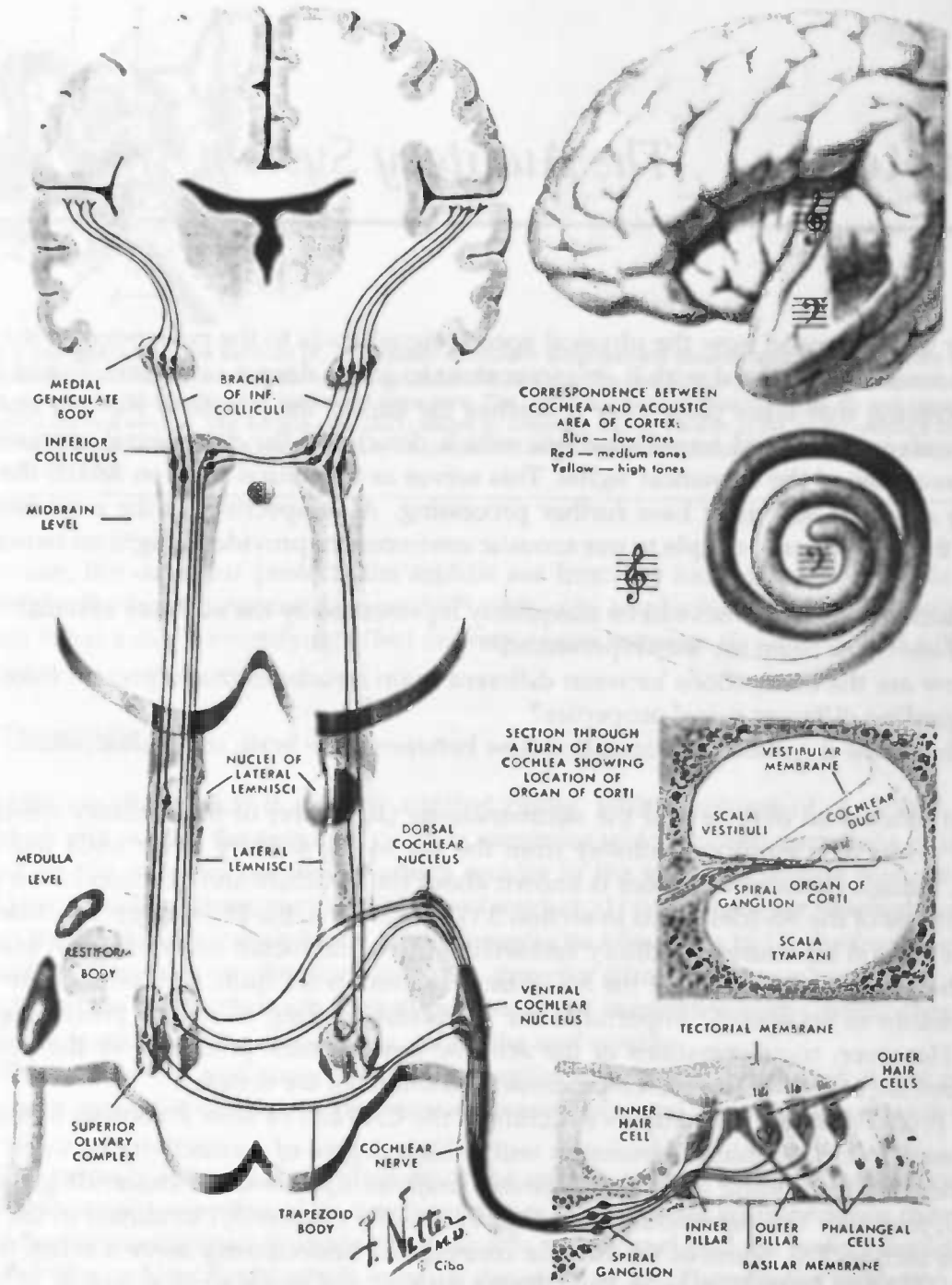


Figure 3.1 The auditory pathway with its nuclei and pattern of connectivity is shown at the left. Also shown in this figure are the correspondence between the representation of frequencies at the cochlea and the acoustic area of the cortex (upper right), a cross-section of the cochlea showing the location of the organ of Corti (middle right), and the organ of Corti (lower right), which will be discussed later in this chapter.

Knowledge of auditory functioning is mainly based upon measurements with simple stimuli. Within the neuroethological approach, emphasis is placed on performing measurements on "auditory specialized" animals (e.g., the mustached bat) by using species-specific behaviorally significant sounds. These measurements provide insights on the upper limits of specialization of the neural processing mechanisms at the level of the auditory cortex, and therefore provide a useful additional source of knowledge on auditory functioning. Some speculations regarding human speech processing from a neuroethological point of view will be presented in section 3.9.

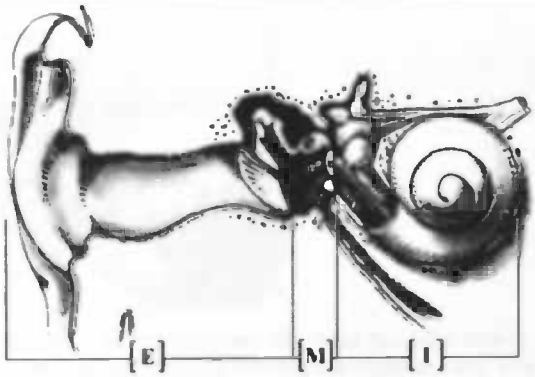


Figure 3.2 Schematic representation of the ear, consisting of the external ear (E), the middle ear (M), and the inner ear (I).

3.1 The ear

The sound waves that enter the ear are compressional waves. They can be characterized by their instantaneous sound pressure $p(t)$. The amplitude of the oscillations in $p(t)$ is relatively very small compared to the barometric pressure P_0. This acoustical signal serves as the input to the peripheral auditory system, the output of which consists of a collection of neural spikes that enter the brain. In short, what happens in-between can be summarized as follows.

The process consists of three stages corresponding to the three gross anatomic structures (i.e., the outer, middle and inner ear) that comprise the peripheral auditory system (see figure 3.2):

1) The sound wave enters the *outer* (or *external*) ear (and travels through the *auditory canal*);
2) Movement of the *tympanic membrane*, the boundary between the outer and *middle ear*, is transferred by the *ossicles* of the middle ear to the *oval window* which is the boundary of the middle and *inner ear*;
3) Movement of the oval window generates a pressure difference in the fluids of the *cochlea*, which, in turn, moves the *basilar membrane* upon which the primary auditory receptors, the *hair cells* are located. The hair cells are mechanosensors converting motion of their hair bundles into neural spikes.

3.1.1 The outer ear

The outer ear consists of the *pinna* and the *auditory canal* (*meatus*).

- The pinna is relatively unimportant, but does modify the incoming sound, particularly at high frequencies, which is of some importance in our ability to localize sounds. Furthermore, the structure of the pinna makes it attenuate certain frequencies more than others. At frequencies below about 1,000 Hz, the auditory meatus has essentially no directional effect.
- The auditory canal is an acoustic tube with a length of about 2.4 cm on average. Sound travels down the auditory canal and is transmitted to the *eardrum*. Here, acoustic energy is transduced to vibrational mechanical energy of the *tympanic membrane* at the middle ear.

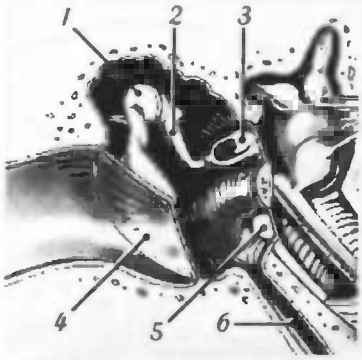


Figure 3.3 The *middle ear*. The *eardrum* (4) or *tympanic membrane* separates the *external auditory meatus* from the middle ear which communicates with the mouth cavity via the *Eustachian tube* (6). The *oval window* (hidden by the stapes foot plate: 3) and the *round window* (5) separate the middle and inner ears. The ossicular chain: *malleus* (1), *incus* (2) and *stapes* (3) links the eardrum to the oval window. The surface ratio (20/1) allows an adequate energy transfer of the sound pressure between the air and the fluids of the inner ear. The middle ear can be considered as an impedance adapter - without it about 98% of energy would be reflected back.

Furthermore, the outer ear protects the middle ear from the hazards of the outside world. Interestingly, the human ear canal resonates[23] such as to emphasize the spectral components of speech: it has a slight amplifying effect on frequencies between about 2,000 and 6,000 Hz.

3.1.2 The middle ear

The middle ear (figure 3.3) is a small airfilled cavity, with a volume of ca. 2 cm³, that is bounded on one side by the *tympanic*, or *drum membrane* which collects variations in air pressure, and on the other by the *cochlea* which resides in the inner ear. Inside this cavity is a complicated linkage of three very small bones (*ossicles*): (1) the *malleus* (or *hammer*) is set into vibration by the tympanic membrane, and transmits its vibrations to (2) the *incus* (or *anvil*), which in turn, transmits its vibrations to (3) the *stapes* (or *stirrup*), where vibrations are finally transmitted to the *cochlea* through a small portion of the membrane surrounding the cochlea in which the footplate of the stapes rests, called the *oval window*.

The middle ear also comprises the *Eustachian tube* which opens and closes periodically to help maintain a constant pressure difference between the middle ear and the ear canal.

The primary functional significance of the middle ear can be understood if one considers the physical properties of its two boundaries. There exists an impedance mismatch[24] between the surrounding air (outside the tympanic membrane) and the watery liquid inside the cochlea (the *cochlear fluid*), which is much denser than air. Therefore, a reduction of this impedance mismatch is required in order to make a transformation of time variations in air pressure to time variations in fluid pressure in the cochlea efficient. Airborne vibrations of the large tympanic membrane are therefore transformed by the following interdependent transformation mechanisms:

[23] Sound waves near the resonant frequency of the auditory canal are partly reflected from the closed end of the canal and reinforce the incoming sound wave of the same frequency, dependent on the length of the canal, at about 3,400 Hz.)

[24] Impedance is a property of the transmission medium. The *characteristic* acoustic impedance (under ideal circumstance, i.e., in a free field) of air is the ratio of pressure and velocity, or the product of the density of the air ρ_0 and the velocity of the waveform c, which is approximately 40 dyne sec/cm³, whereas the characteristic acoustic impedance of the cochlear fluid is about 161,000 dyne sec/cm³ (a factor 4,000 higher). [A dyne = a unit of force: the force needed to accelerate 1 g 1 cm/sec².]

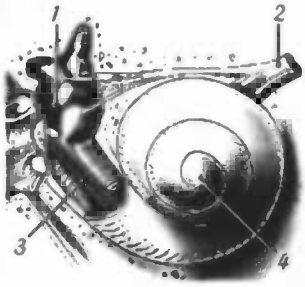


Figure 3.4 The *inner ear*. The bone has been broken to visualize the vestibule (1), the VIIIth nerve (2), and the basal portion of the cochlear duct (3), housing the organ of Corti. The rest of the cochlea (4) is covered by the bony capsule. The VIIIth nerve is formed by the vestibular and cochlear nerves which merge before entering the brain.

(1) By changing the effective area to which the force is applied the pressure per unit area increases while the velocity decreases.
(2) To a smaller extent, the ossicles acts as small levers. According to the lever principle a small force applied to a long lever exerts a larger force, acting through a smaller distance, at the other end of the lever.
(3) And more subtle, the tympanic membrane itself, by virtue of its curvature acts as a mechanical transformer.

In some clinical cases, where the ossicle chain is broken, the acoustic waves in air are directly transmitted to the cochlear window without the mechanical advantage of the middle ear, thereby suffering a considerable loss in intensity (between 20 and 30 dB).

Finally, there are two important muscles in the middle ear, attached to the ossicles:

(1) the *tensor tympani* which is attached to the malleus, and
(2) the *stapedius* which runs from the posterior wall of the middle ear to the neck of the stapes.

Activation of these muscles protects the ear from very loud sounds by changing the mode of vibration of the ossicular chain. The amount of protection is about 20dB but may be greater for brief periods. This protective system is put in operation by an *acoustic reflex* whose latency depends on the intensity of the sound but is approximately 60–120 msec. Consequently, it affords no protection to impulsive sounds (like for example, gun shots).

3.1.3 The inner ear

The inner ear (figure 3.4) consists of the *cochlea* which is a small coiled, fluid-filled tube with a length of approximately 35 mm. It is embedded in the extremely hard *temporal bone*. The cochlea is specialized for transforming mechanical energy into impulses in the nervous system.

In cross section, two very thin membranes divide the tube near the middle, thereby creating three different compartments (or *scalae*). The largest one is the *scala vestibuli*, the next smaller one is the *scala tympani*. These are respectively connected to the middle ear by a flexible membrane, the *oval window*, to which the stapes is attached, and the *round window*, a membrane-covered opening below the oval window. They both contain the same fluid (i.e., *perilymph*), and are connected at the upper end of the cochlea by a small opening near the *apex* called the *helicotrema* which permits the fluid to "communicate" between the two compartments. Separating these two rather large compartments and running almost the entire length of the cochlear tube is the *scala media*, which is separated from scala vestibuli by *Reissner's membrane*, which seems to fulfil no mechanical purpose. Between scala media and

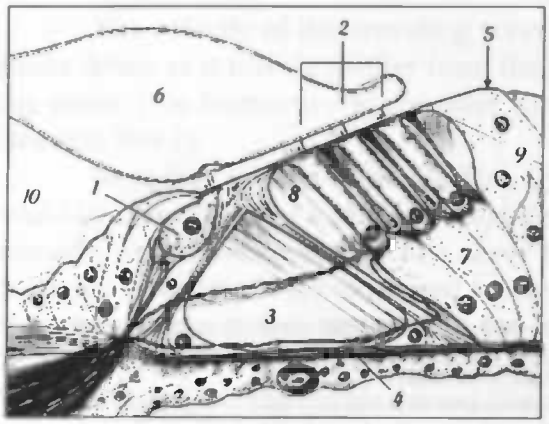


Figure 3.5 *Organ of Corti.* In this schematic drawing from a transverse section in the basal turn of a mammalian cochlea, the two types of sensory cells: *inner* (IHC: 1) and *outer hair cells* (OHCs: 2) are seen on both sides of the *tunnel of Corti* (3) which is limited by the 2 pillars. The *tectorial membrane* (6), floating in the *endolymph* covers the hair cell stereocilia embedding tips of OHC tallest *stereocilia*. The IHC is surrounded by support cells while the OHC is only firmly "seated" on Deiters' cells (7), its lateral membrane being in direct contact with the corticolymph (almost identical to the *perilymph*) which fills in the tunnel (3) and the spaces of Nuel (8).The cuticular plate of hair cells, together with the head of pillars, the phalangeal processes of Deiters' cells and the apical membrane of other support cells, such as Hensen's cells (9), form the reticular lamina (5) sealing the endolymphatic compartment. Breaking through *the basilar membrane* (4), *nerve fibers* reach or leave the organ of Corti .

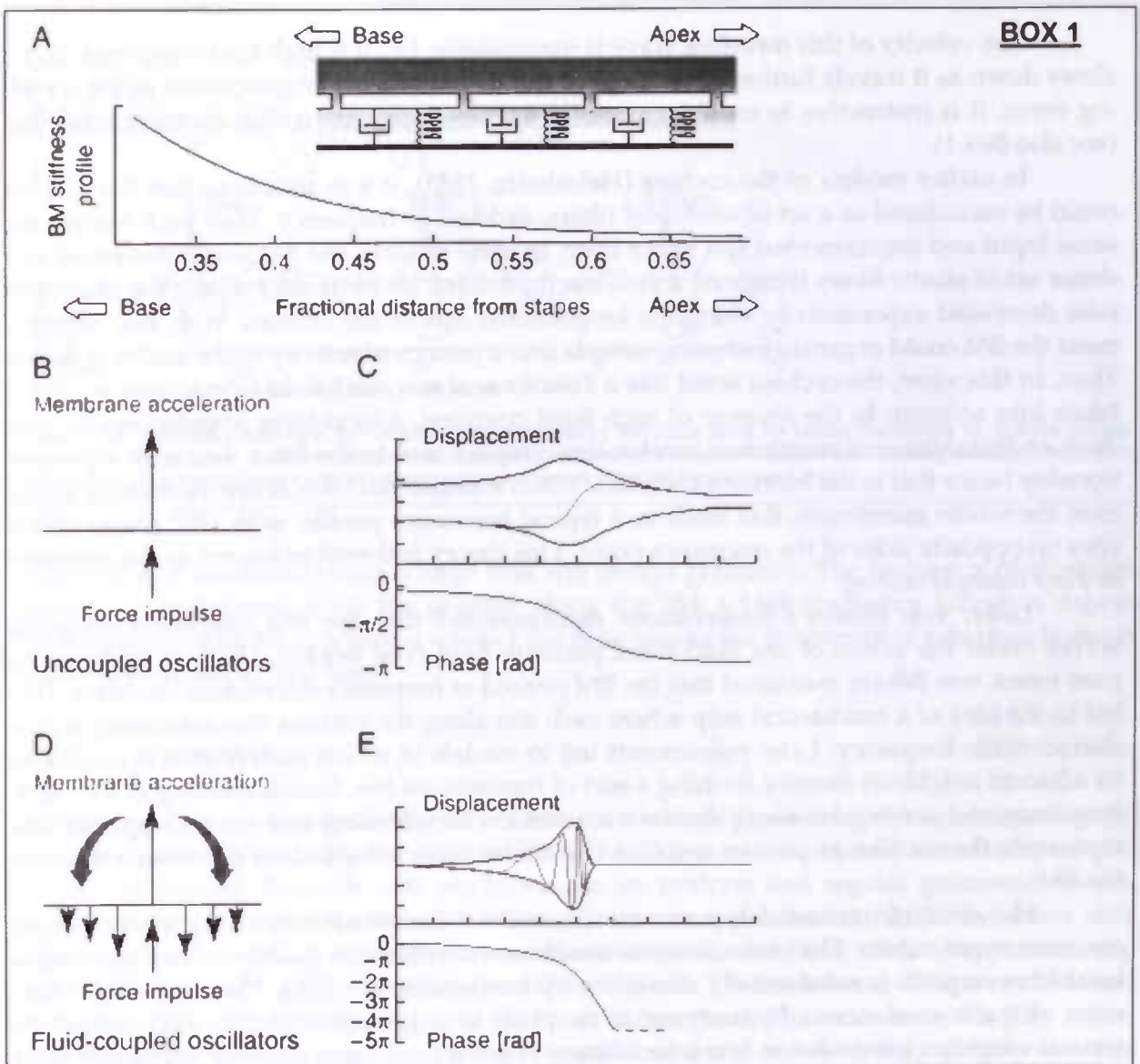
scala tympani a complicated arrangement of cells and membranes, which forms the *organ of Corti*, is found, with on the side of the scala tympani, the *basilar membrane*. This membrane is assumed to contribute the major part of the mechanical properties of the organ of Corti. A schematized representation of the organ of Corti is depicted in figure 3.5.

Sensory receptors, known as *hair cells*, reside on the basilar membrane and project toward the overlying *tectorial membrane* supported by a framework of *pillar cells*. These hair cells are the final transducing medium. The motion of the BM, resulting from sound vibrations, precedes the major transformation of the acoustic wave from a mechanical disturbance to a pattern of neural excitation. An understanding of the exact form of this motion is essential to any auditory theory.

3.1.3.1 *The basilar membrane*

The motion of the oval window sets the fluid in the cochlea in motion. As the fluids and the surrounding bone are essentially incompressible, the fluid displaced at the oval window by the movement of the stapes must be equalized, which occurs at the round window. The motion is transmitted to the *basilar membrane* (BM), a fiberous tissue which spans about half the width of the cochlea. The sound wave is a pressure difference in the fluids of the cochlea which may be treated as acting instantaneously over the entire membrane. [25] In response to these forces the BM starts to move: a wave motion results that travels from the stapes, where the area of cochlear duct is largest, down the membrane toward the helicotrema, where the area is smallest. The pattern of motion takes some time to develop, the membrane is first displaced near the *base*, and the displacement then moves to the *apex*. This pattern has been described as a *traveling wave* (von Békésy, 1960). The waves in the BM disappear in the vicinity of the helicotrema, so there is no prominent reflection back toward the stapes.

[25] As the fluids in the inner ear are considered to be similar in composition to sea water where the velocity of sound is approximately 160,000 cm/sec, the wavelength of a high frequency of for example 10 kHz is approximately 16 cm. The cochlear duct is about 35 mm long and therefore, even for a high frequency, only 1/5th of the wavelength of the sound. Therefore, at any instant, essentially the same pressure extends from one end of the duct to the other: nearly the same physical force is applied throughout its length as the stapes vibrates and creates a sound field in the fluids of the scala vestibuli. Thus, sound may be treated as a pressure difference between scala vestibuli and scala tympani which, at any instant in time, is uniform over the entire extent of the membrane.



The basilar membrane of the mammalian cochlea is an elastic strip tensioned across the fluid-filled cochlear duct. The membrane, whose different portions interact through the fluid, is the supporting medium for "traveling waves" that form when sound is transmitted through the middle ear. For the sake of physical clarity, the basilar membrane can be conveniently pictured as a bank of spring-mounted micro-pistons (modules) immersed in an inviscid fluid (Fig. A, inset). Each piston's stiffness is an exponentially decreasing function of the distance from the stapes (Fig. A).

To explain the formation of traveling waves, as described by von Békésy (1960), how such an array of modular oscillators interact instantly with each other through the fluid must be analyzed. The instantaneous character of the fluid-mediated coupling amongst the modules is a consequence of the negligible delay (about 10 μs) when the pressure field propagates within the cochlear duct compared with the basilar membrane oscillation periods (typically ranging from 50 ms to 50 μs). As can be inferred from direct observation of basilar membrane motion (Johnston et al., 1986; Robles et al., 1986), in normal acoustic conditions the minimum wave length of the traveling wave spans five to ten modules. For the mathematics it is therefore more satisfactory to treat the micro-piston system as a continuum.

In the absence of fluid coupling, a local-force impulse would produce an instantaneous membrane acceleration proportional to the force, with the proportionality factor being the inverse local mass (Fig. B). Correspondingly, under the action of a sinusoidal force synchronously acting over the whole membrane, the response would be a typical resonance profile (Fig. C), with 180° phase difference on opposite sides of the resonance point. This "dry cochlea" model was originally proposed by Helmholtz (1885) and influenced cochlear physiology for over a century.

Hydrodynamic coupling alters substantially this acceleration pattern. The instantaneous membrane acceleration elicited by a local-force impulse applied to the membrane at rest (as above) displays the profile in Fig. D. Fluid pressure spreads instantaneously from the force application site and pushes adjacent membrane segments in the direction opposite to the force impulse (curved blue arrows). Thus, in general, any local basilar membrane oscillation generates forces that tend to drive flanking modules to swing with opposite phases. However, due to the exponentially graded membrane stiffness, the effect is different at opposite sides, the response being smaller at the stiffer (more basal) side, so that the semi wavelength of the oscillation decreases from base to apex (Fig. E). Under the action of a sinusoidal input at the stapes, this dynamic effect results in a monotonically increasing phase delay of the local oscillations versus distance from stapes. This asymmetry bears important consequences on wave dynamics as the "apparent mass" of the fluid locally involved in the oscillation decreases from base to apex far more rapidly than all other graded quantities, resulting in the characteristic shrinking of the traveling wavelength at a frequency-dependent critical point. The wave amplitude profile tends to elevate markedly in the proximity of this point, beyond which the membrane motion, as well as that of the fluid, undergoes a steep fall. The peak height of the traveling wave would increase without limit if the intrinsic viscosity fell to zero (Mammano and Nohili, 1993). [From Nobili et al., 1993]

The velocity of this traveling wave is nonuniform, i.e., it is high near the stapes, and it slows down as it travels further from the stapes. To understand the generation of the traveling wave, it is instructive to consider some of the developments within cochlear modeling (see also Box 1).

In earlier models of the cochlea (Helmholtz, 1885), it was proposed that the cochlea could be considered as a set of *uncoupled* filters, ordered in frequency. They each receive the same input and are connected to a nerve fiber. In these models, the BM was represented as a dense set of elastic fibers tensioned across the fluid-filled cochlear duct where the fiber stiffness decreased exponentially along the longitudinal axis of the cochlea. With this arrangement the BM could organize incoming sounds into a pattern of activity in the auditory nerve. Thus, in this view, the cochlea acted like a Fourier analyzer, and fluid interactions were not taken into account. In the absence of such fluid coupling, a local-force impulse would produce an instantaneous membrane acceleration proportional to the force, i.e., with a proportionality factor that is the inverse local mass. When a sinusoidal force is synchronously acting over the whole membrane, this leads to a typical resonance profile, with 180° phase difference on opposite sides of the resonance point. This theory is therefore known as the *resonance* or *place theory of hearing*.

Later, von Békésy's observations demonstrated that the BM supported traveling waves under the action of the fluid-filled pressure field (von Békésy, 1960). In response to pure tones, von Békésy measured that the BM peaked at frequency-dependent locations. This led to the idea of a mechanical map where each site along the cochlea was associated with a characteristic frequency. Later refinements led to models in which each filter was *coupled* to its adjacent neighbors thereby forming a sort of *transmission line*. Sound entering at the high-frequency end propagates along such a transmission line, leading to a neural response that represents the oscillation pattern associated with frequency-dependent traveling waves on the BM.

However, current models are even more refined, because the dynamics of the cochlea are even more subtle. The instantaneous membrane acceleration pattern when applying a local force impulse is substantially altered by hydrodynamic coupling. The long-range character of the instantaneous hydrodynamic coupling is only approximately represented by nearest-neighbor transmission-line interactions. When a local-force impulse is applied, fluid pressure spreads instantaneously from the force application site and pushes adjacent membrane segments to the direction opposite to the force impulse. Thus, in general, any local BM oscillation generates forces that tend to drive flanking modules to swing with opposite phases. This effect is different at opposite sides due to the exponentially graded membrane stiffness. The response is therefore smaller at the stiffer (more basal) side, so that the semi wavelength of the oscillation decreases from base to apex.

So, when the input at the stapes is a continuous sinusoidal vibration, this dynamic effect results in a monotonically decreasing phase delay of the local oscillations versus distance from stapes. This asymmetry has important consequences on wave dynamics. The "apparent mass" of the fluid that is locally involved in the oscillation decreases from base to apex far more rapidly than all other graded quantities, resulting in the characteristic shrinking of the traveling wavelength at a frequency-dependent critical point. In the proximity of this point, the wave amplitude profile tends to elevate markedly. Beyond this point, the membrane motion as well as that of the fluid, attenuates rapidly. Each point along the BM has its own characteristic frequency (CF), so the BM transforms frequencies present in the stimulus input into places along the BM (see figure 3.6).

A very brief, impulsive sound leads to a pattern of vibration along the entire length of the BM, as by definition, an impulse contains all frequencies with approximately equal amplitudes (resulting in a flat spectrum). The displacement response travels again from the stapes to the helicotrema, but the activity is less peaked. At any local region, the generated

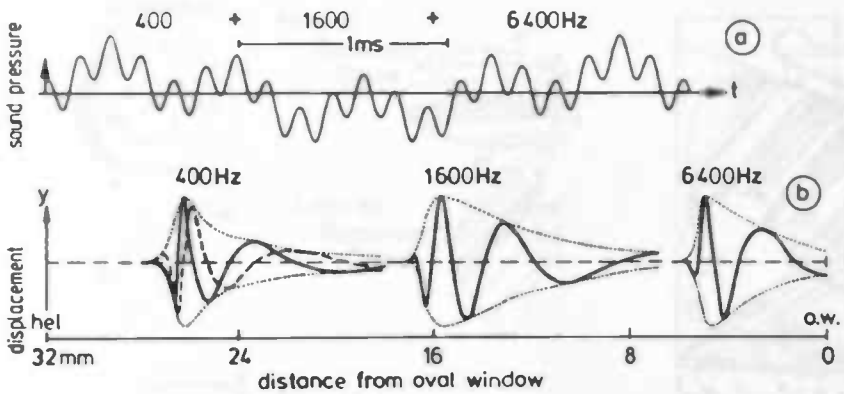


Figure 3.6 Schematic drawing of the transformation of frequency into place along the basilar membrane. In (a) three simultaneously presented tones of different frequencies expressed as compound time function produce travelling waves (b), that reach their maximum at three different places corresponding to the characteristic frequencies (CFs).

response as a function of time is large first and decays gradually. The frequency of vibration of this decay depends upon the position along the BM: a high-frequency vibration decays fairly quickly with time, whereas toward the helicotrema the frequency of vibration is much lower and the decay rate less.

3.1.3.2 *Hair cells*

It has already been mentioned that the vibration of the BM generates the mechanical forces that stimulate the hair cells. In what follows, it will be considered how these receptor elements are excited. The hair cells are located in an intricate and regular geometric pattern along the length of the basilar membrane. There are approximately 3,500 *inner hair cells* (IHCs) that are located as a single row on the inner side, and three rows of approximately 20,000 *outer hair cells* (OHCs) further away from the axis. The anatomical differences between these two types of hair cells suggest differences in functions. The inner hair cells play a central role in the transmission of signals to the auditory nerve to which they directly synapse (see also section 3.2). The first temporal derivative of the BM displacement, the velocity, is often assumed to be the *effective* stimulus driving the inner hair cells.[26] The outer hair cells are thought to provide *mechanical amplification* to the BM motion. Furthermore, the OHCs are the target of descending fibers from higher centers of the auditory pathway, but the exact implication of this has not been very well understood yet.

Incoming sound vibrations cause the basilar membrane to move up and down. As a result the organ of Corti (figure 3.5) is displaced to the plane of the BM so that the position of hair cells relative to the tectorial membrane is altered. This causes a relative motion between the basilar and tectorial membrane, resulting in a *shearing force* on the outer hair cells that are particularly sensitive to forces applied in certain (longitudinal as well as radial) directions. The small stiff *cilia* (or *hairs*) on top of the hair cells are thus subjected to a mechanical force, provided by the tectorial membrane into which the tallest stereocilla of the OHCs are inserted. This leads to the opening of ion channels in the cilia and results in the generation of a depolarization of the hair cell. It has been shown that OHCs can produce length changes when placed in an electric field mimicking this depolarization, as well as due to mechanical stimulation of the hair bundle. The elongation and shortening of the OHCs occurs at the same rate, so the OHCs vibrate at the rate of stimulation. This OHC motility forms the *active*

[26] Notice that this results in a relative amplification of high-frequency components, since their velocity is faster than for low-frequency components. However, in combination with the larger source amplitude of the lower compared to the higher frequency components (see section 2.1), this leads to a flattening of the spectrum.

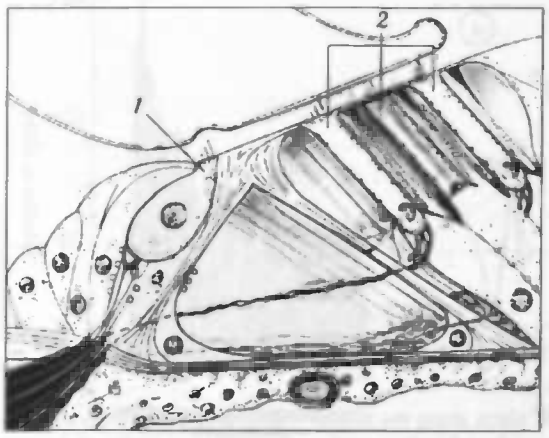


Figure 3.7 Innervation pattern of the inner and outer hair cells. (a) Both hair cells are innervated by specific afferent and efferent systems: the radial afferent (blue) and the lateral efferent (pink) for the IHC (1), the spiral afferent (green) and the medial efferent (red) for the OHC (2)..

mechanism of the organ of Corti, and has been hypothesized to have a direct influence on the BM displacement, in the sense that it can *enhance* the frequency tuning and selectivity.

The OHC response does not increase linearly with increasing intensity, but it shows signs of saturation, even around 40-60 dB (the typical range for speech signals). At input levels that correspond to more than 10 dB SPL, the amplification characteristic becomes more and more saturated. This results in a *compressive nonlinearity* and has an influence on the frequency-selectivity, i.e., at higher levels the frequency selectivity is reduced due to a broader tuning. [27]

3.2 The eighth nerve

Inner hair cell responses are communicated to the nervous system via the *auditory nerve*. Stimulation of an inner hair cell in the cochlea causes depolarization[28] inside that cell thereby releasing a neurotransmitter that causes a depolarization in the nerve fiber adjacent to the cell. These receptor cells reside in the *spiral ganglion*, and as a result neurotransmitters are released. There are approximately 30,000 fibers emanating from the ganglion to innervate the inner hair cells in a specific pattern (approximately 10 neurons per IHC). The innervation pattern of the IHCs and OHCs is depicted in figure 3.7. The innervating fibers form the *afferent nerve*. *Efferent* fibers from the spiral ganglion cells project to higher auditory structures and form the *eighth nerve*.

[27] This explains for a part the discrepancy that was thought to exist between the form of the traveling wave as observed by von Békésy (1960) and the sharpness of *tuning curves* of auditory nerve fibers (see also section 3.2), the latter being much sharper than the former. Von Békésy's measurements were performed with high sound levels. Furthermore, these measurements were performed on a passive BM, i.e., in dead animals. Note also that the enhancement of frequency selectivity through suppression of adjacent frequencies is explained as a *mechanical effect* that is equivalent to *lateral inhibition* in neural structures. Earlier theories explained the observed discrepancy in terms of such inhibitory neural processes (Moore, 1977). Later, explicit reference to certain neural processes was avoided by speaking in terms of neural *suppression*. With the current insights in cochlear mechanics, there seems to be no true discrepancy, because of the presence of the active mechanisms in the cochlea in the intact cochlea.

[28] When the membrane of a neuron is excited, its permeability for some ions changes abruptly. As a result the potential of the inside of the membrane (this resting potential is normally approximately 100 mV negative with respect to the outside) quickly becomes less negative and may even become positive for a short period of time. Within about a millisecond, regeneration starts and the ion balance is restored. The produced discharge (action potential) travels along the cell membrane and its branches, in particular the axons.

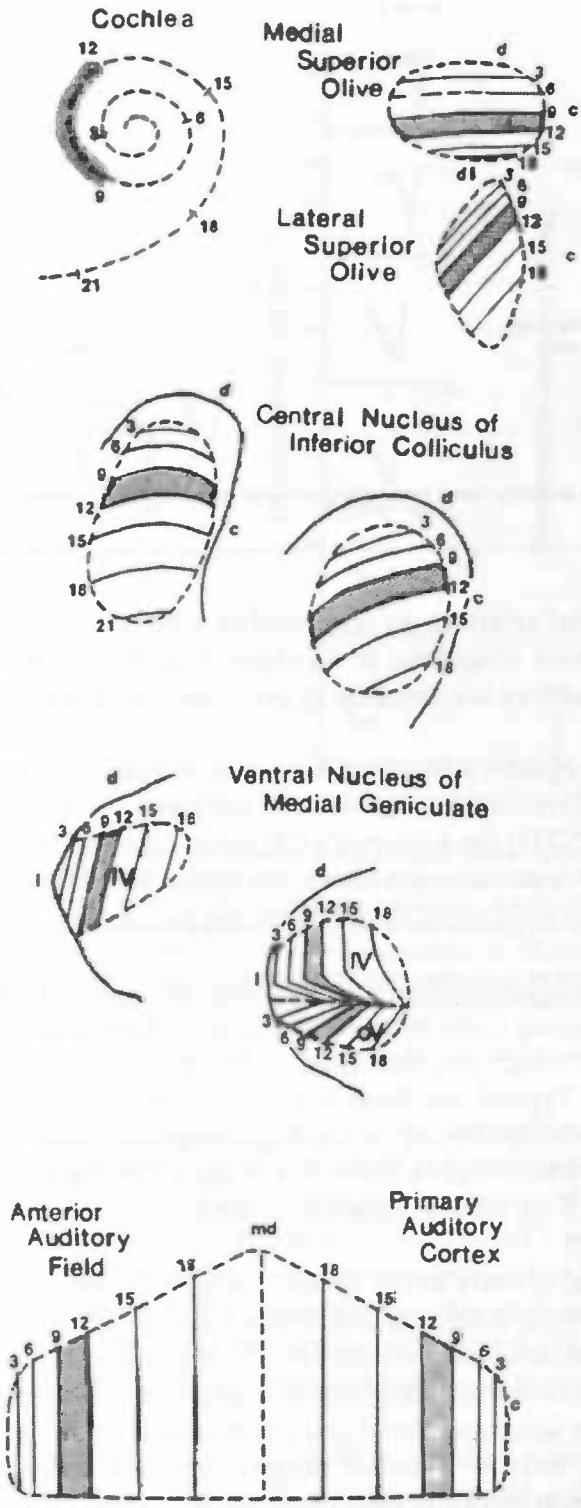


Figure 3.8 Schematic representation of the cochleotopic/tonotopic organization within each of the major nuclei of the auditory nervous system. The bands indicate 3 mm section of the cochlear sensory epithelium with the 9-12 mm section shaded. [From Moore, 1995].

The nature of the representation of acoustic information in the eighth nerve is important, because it forms the only source of information available to higher levels of auditory processing.

It appears that the distribution of frequency that occurs peripherally in the cochlea is preserved in the AN fibers, i.e., the frequency to which a fiber responds best, its best frequency (BF), is determined by that part of the BM where it innervates an inner hair cell.

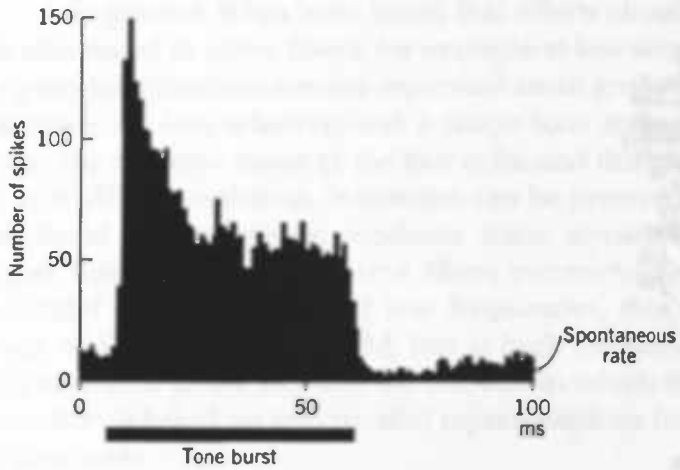


Figure 3.9 *Post-Stimulus Histogram* (PSTH) from an afferent auditory nerve innervating an inner hair cell, showing adaptation effects in AN fibers.

Nerve fibers also tend to maintain their spatial relations to one another which results in a systematic arrangement of frequency responses according to location. This has been observed in many centers of the brain where auditory information is processed, and is known as *tonotopic organization* (see figure 3.8).

All fibers in the AN show adaptation effects: a maximal response at stimulus onset, which gradually decreases and reaches a steady-state after som tens of milliseconds. This can be shown in a *Post-Stimulus Time Histogram* (PSTH) for a neuron's CF, which is a representation of the discharge rate as a function of time from stimulus onset (see figure 3.9). The auditory nerve fibers in the eighth nerve differ from each other in the following properties:

- *Frequency response range*: They show *frequency selectivity* which is often illustrated by the tuning curves of single nerve fibers. A *tuning curve* represents the cell's threshold as a function of frequency. The frequency at which the threshold of the fiber is lowest is called the characteristic frequency (CF). Typical for these tuning curves is that, on a logarithmic frequency scale, they are generally steeper on the high-frequency side than on the low-frequency side, especially for fibers with a low CF. For high CFs the tuning curves are more symmetrical (see figure 3.10, left). Furthermore, tuning curves are related to the rate of discharge of a neuron.[29]
- *Amount of spontaneous activity* (background activity in the absence of sound stimulation) and *maximum fire rate*. The *general* shape of the discharge curve as a function of intensity is characteristic for most AN fibers. Often *saturation* occurs which entails that above a certain sound level there is no longer an increase in firing rate in response to an increase in sound level. For example, for the cat the saturation level lies within about 40 dB of the hearing threshold, so there is a relatively narrow dynamic range compared to the dynamic range of 100 dB to which the human auditory system is sensitive.

[29] For example, the *iso-rate contours*, determined by the intensity of stimulation required to produce a predetermined firing rate in the neuron as a function of frequency, are generally similar in shape to tuning curves. In contrast, *iso-intensity contours* generally vary in shape according to the sound level chosen, and differ considerably from tuning curves. As the discharge rate varies nonlinearly with intensity, iso-intensity contours are more difficult to interpret. Also, for some fibers the frequency that produces a maximal firing rate varies as a function of intensity level. This poses problems for some theories of pitch perception.

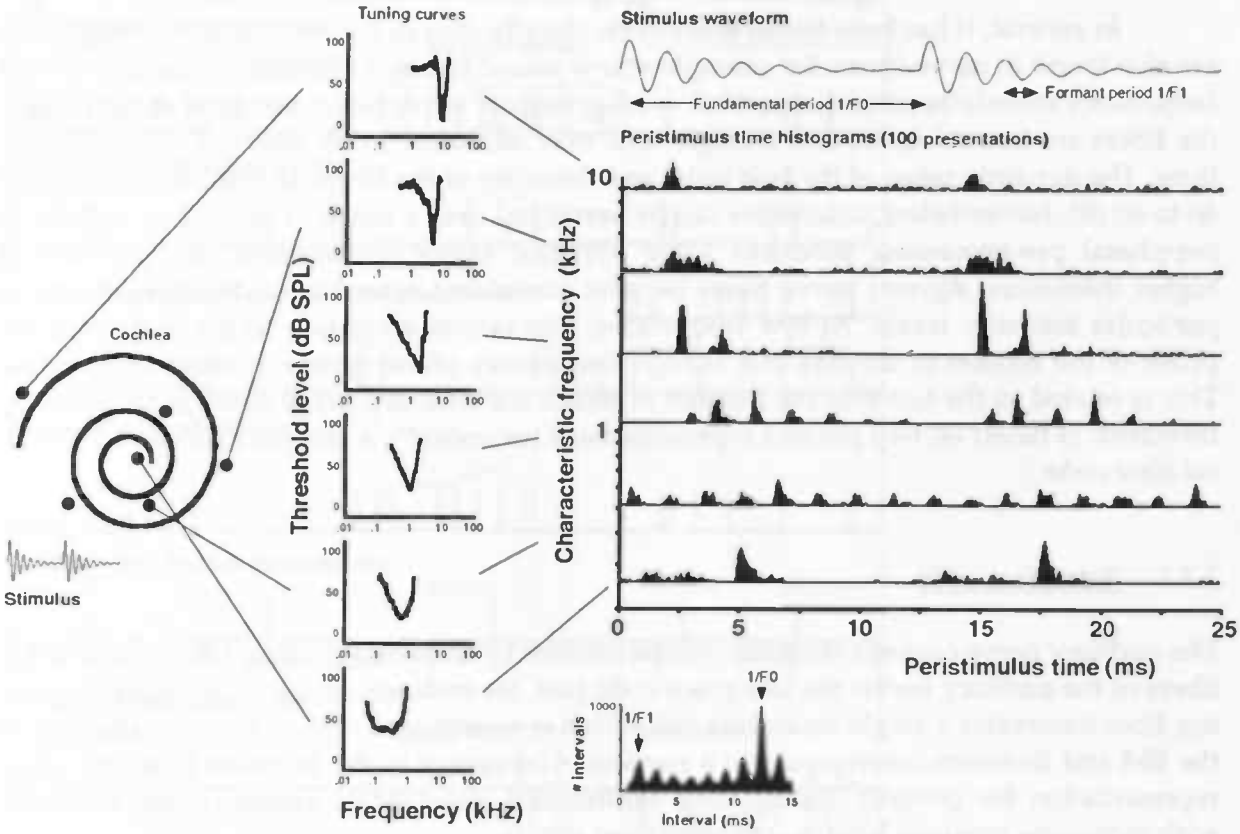


Figure 3.10 Temporal patterns in the auditory nerve. Tuning curves and peristimulus time (PST) histograms for six representative auditory nerve fibres with different characteristic frequencies (CFs). The stimulus is a periodic wave with a fundamental period of 12.5 ms (F0 = 80 Hz) and one resonance period at 1.6 ms (F1 = 640 Hz), a "one-formant vowel". Left: relation between the representation of frequency at places at the cochlea and the tuning curves of neurons in the AN with a certain CF. Right: all-order interspike interval histogram for a population of 79 auditory nerve fibres, showing different time patterns in the different frequency channels. Several discharge periodicities are related to 1/F0 and 1/F1. In the population interval distribution (bottom center), the most common interspike interval corresponds with the perceived pitch period of the stimulus (12.5 ms). [From Cariani, 1995].

- *Phase-locking capability*: Neural firings tend to occur at a particular phase of the stimulating waveform, so that there is a temporal regularity in the firing pattern of a neuron in response to a (quasi)periodic stimulus. If the duration of the stimulus is longer than the duration of an action potential (about 1 ms), then the fibers can phase-lock to the stimulus. Since individual neurons can no longer fire with the period of the stimulus, groups of phase-locked neurons together represent the stimulus. In the squirrel monkey, this capability has been observed up to 5 kHz (Rose, 1967). It is generally believed that the upper limit for phase-locking lies at about 4-5 kHz.[30]

These properties are often determined by recordings with micro-electrodes from single auditory neurons which, off course, does not tell us anything about the pattern of neural responses over different auditory neurons, or about possible interactions or mutual dependencies in the firing patterns of different auditory neurons. (*Effective*) *excitation patterns* are often determined indirectly through extrapolation on the basis of these single neuron properties (Moore, 1977), but due to possible interactions this may not be a valid procedure.

[30] However, this upper limit is *not* determined by the refractory periods of neurones or by their maximum fire rates (which is sometimes claimed), but by the precision with which the initiation of a nerve impulse is linked to a particular phase of the stimulus. There is a certain *"jitter"* in the exact instant of initiation of a nerve impulse which at high frequencies becomes comparable with the period of the waveform, so that above a certain frequency the spikes will be "smeared out" over the whole period of the waveform, instead of occurring primarily at a particular phase. This smearing is responsible for the loss of phase-locking above 4-5 kHz (Green, 1976).

In general, it has been found that effects already seen in the peripheral pre-processing are also found in nerve fibers, for example at low sound levels, a stimulus containing several frequencies stimulates several separated small groups of nerve fibers, whereas at high levels, the fibers are less selective, and a single tone may stimulate many fibers at different locations. The dynamic range of the hair cells, and therefore of the fibers, is assumed to be about 40 to 60 dB. Nevertheless, intensities can be perceived over a range of more than 100 dB, as peripheral pre-processing produces some dynamic range compression, and because at higher intensities, *different* nerve fibers become stimulated according to their sensitivity to particular intensity levels. At low frequencies, this occurs according to the instantaneous phase of the motion of the BM, but at high frequencies phase synchronization disappears. This is related to the fact that the manner in which the auditory nerve conveys stimulus information is based on two parallel representations (or codes[31]), a *rate-place code* and a *temporal-place* code.

3.2.1 Rate-place code

The auditory nerve conveys stimulus *spectral* content by the average firing rate in each of the fibers of the auditory nerve: the rate-place code (see, for instance, figure 3.11b). Each ascending fiber innervates a single inner hair cell which is sensitive to motion in a specific area of the BM and therefore corresponds to a particular frequency in the stimulus. The rate-place representation for periodic stimuli (like synthesized steady-state vowels) often degrades with increasing stimulus level due to saturation effects.

However, for fibers with low spontaneous firing rates which have higher thresholds (Liberman, 1978), significant formant structure is retained in the rate profiles over a much broader range of stimulus levels, which is sufficient for discrimination. It is also suggested that efferent connections from higher processing centers can dramatically improve upon rate-place representations and that such efferent feedback to the cochlea may serve a central role in vowel discrimination in high-noise environments (Sachs et al., 1988a,1988b).

3.2.2 Temporal-place code

The temporal-place code provides a representation of the *temporal* properties of the signal. For example, as a result of the capability of firing in synchrony with the stimulus, activity is directly correlated with time-varying amplitude components of the signal. As depicted in figure 3.11, other temporal coding schemes are also possible. In general, any property of a spike train which covaries with some property of a stimulus can be used to transmit information about that stimulus. The many different coding schemes range from simple interspike interval codes to more complex temporal pattern codes, latency, and history-dependent codes (Cariani, 1995). In particular, it has been found that different auditory nerve fibers are capable of locking/synchronizing to different harmonics of stimuli (Sachs et al., 1988).

[31] A distinction can be made between *signs* and *codes*, which is really the distinction between an observed regularity of nature and an observed regularity which is involved in some identifiable functional role. Codes are functional organizations that actually utilize a particular set of signs to effect a perceptual discrimination. This distinction is analoguous to the distinction between *data* and *information*.

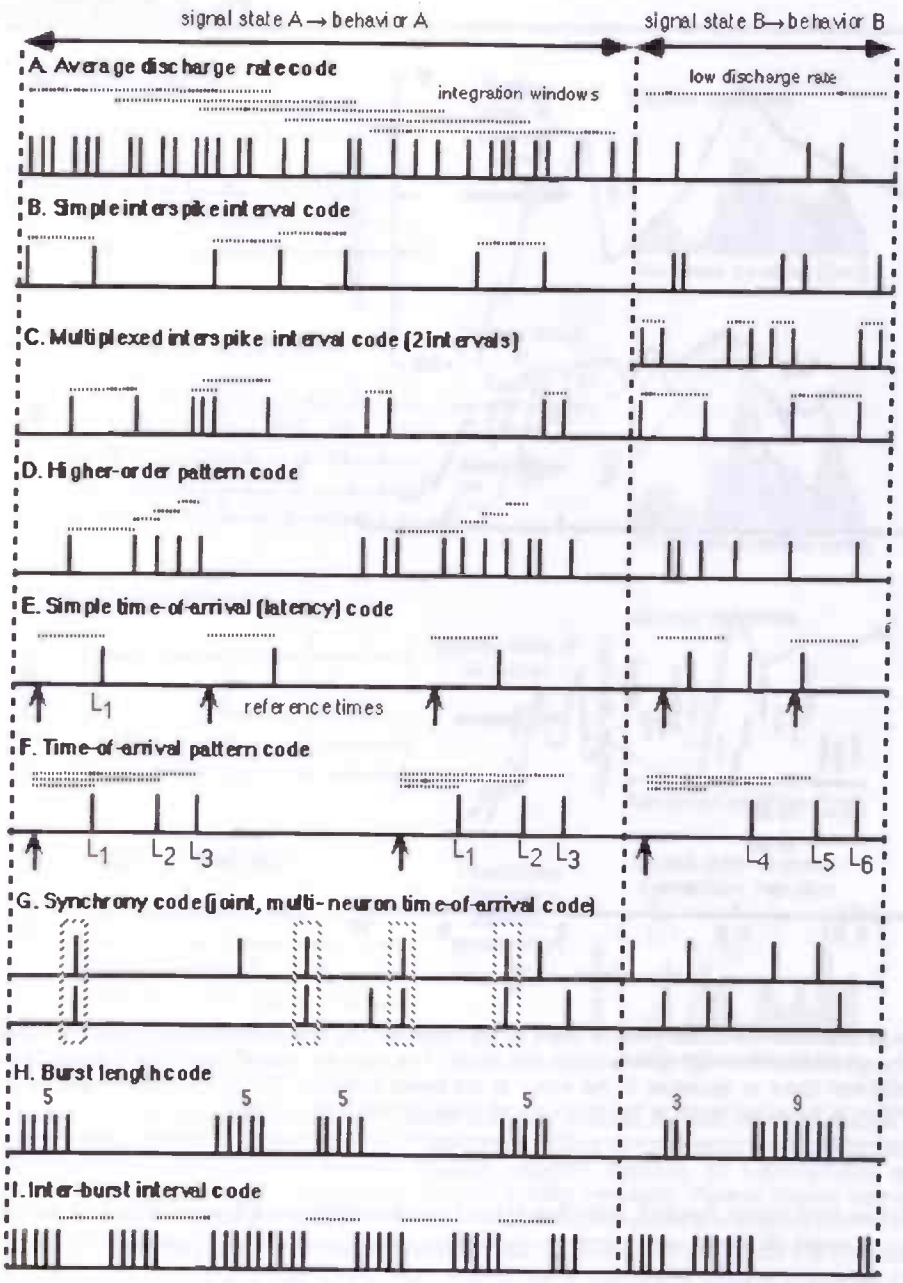


Figure 3.11 Some possible neural pulse codes. These different coding schemes in the time domain range from simple interspike interval codes to more complex temporal pattern codes, latency and history-dependent codes. The spike trains on the left contain patterns that would be recognized as encoded "signals" while those on the right for the most part are examples of patterns that would either be interpreted as different signals or as the absence of signals. [From Cariani, 1995].

Also, based on synchronization measures, neural ensembles appear to be sensitive to formants: formant trajectories (related to *timbre*) are well represented by the *Average Localized Synchronized Rate* (ALSR)[32] which is a way of obtaining information by combining data from many fibers based on the same stimulus (Sachs et al., 1988), and providing a more accurate representation than can be obtained from average discharge rates (see figure 3.12).

[32] *Average Localized Synchronized Rate* (ALSR): derived from the *period histogram* which estimates the instantaneous rate of discharge through one period of a (periodic) stimulus. An FT of the period histogram has amplitude maxima at the stimulus components to which the fiber activity is synchronized. The ALSR is computed by averaging the magnitudes of the FT components corresponding to the stimulus frequency for all fibers whose CF is within one-half to one-quarter octave of the stimulus frequency. This "temporal" measure of synchronization is only computed for those fibers *with a specific CF* corresponding to a "place" on the BM.

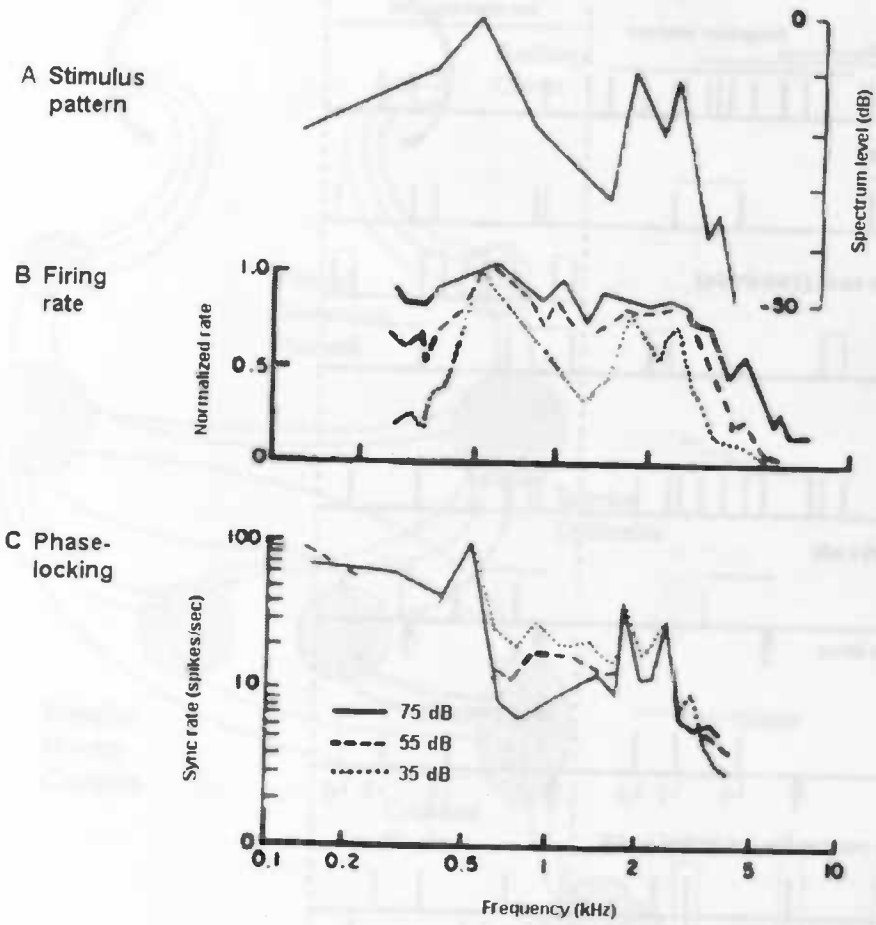


Figure 3.12 Responses of a large population of auditory-nerve fibers to the vowel /ɛ/. (A) The spectral envelope of the vowel. (B) The distribution of mean discharge rates of the high spontaneous rate fibres. The lines are a moving-window average of the mean rates of a population of individual fibers in response to the vowel at the levels indicated. (C) The ALSR function for the same population of fibres in response to the same vowel at the three sound levels. [From Moore, 1995].

Such a sensitivity is not only found for the first two formants, F1 and F2, but some higher formants, for example F3 (Sachs et al., 1988), are also represented at a peripheral level. It has been mentioned in chapter 2 that for vowel discrimination the lowest two formants are the most important, or even sufficient. Though this does seem to hold true for the *full* system, i.e., when observing subject's responses, at a peripheral level, information about higher formants could, in principle, be available.[33] Formant transitions are also well represented by the firing patterns of individual neurons as reflected in *Post-Stimulus Time Histograms* (Delgutte, 1997).

The temporal-place representation is capable of retaining detailed spectral information at high stimulus amplitudes. It therefore appears to be more robust for periodic stimuli (such as vowels and voiced stop consonants) than the rate-place representation. But, it is probably less adequate for unvoiced fricatives and other unvoiced speech sounds for which the energy is often concentrated at higher frequency ranges and therefore, less accurate phase-locking capabilities (Sachs, 1984).

[33] It should be noted however, that the observed AN fiber responses to F3 have been found in *cats* in response to vowel stimuli. It is possible that, though the information about higher formants might be available at a peripheral level, this information is not used (or at least, less important) at the highest levels of processing in humans *when perceiving vowels*, because it is not needed for vowel discrimination. This might explain the relative ignorance to F3 or F4 in vowel discrimination tasks. For consonant discrimination, on the other hand, F3 is important.

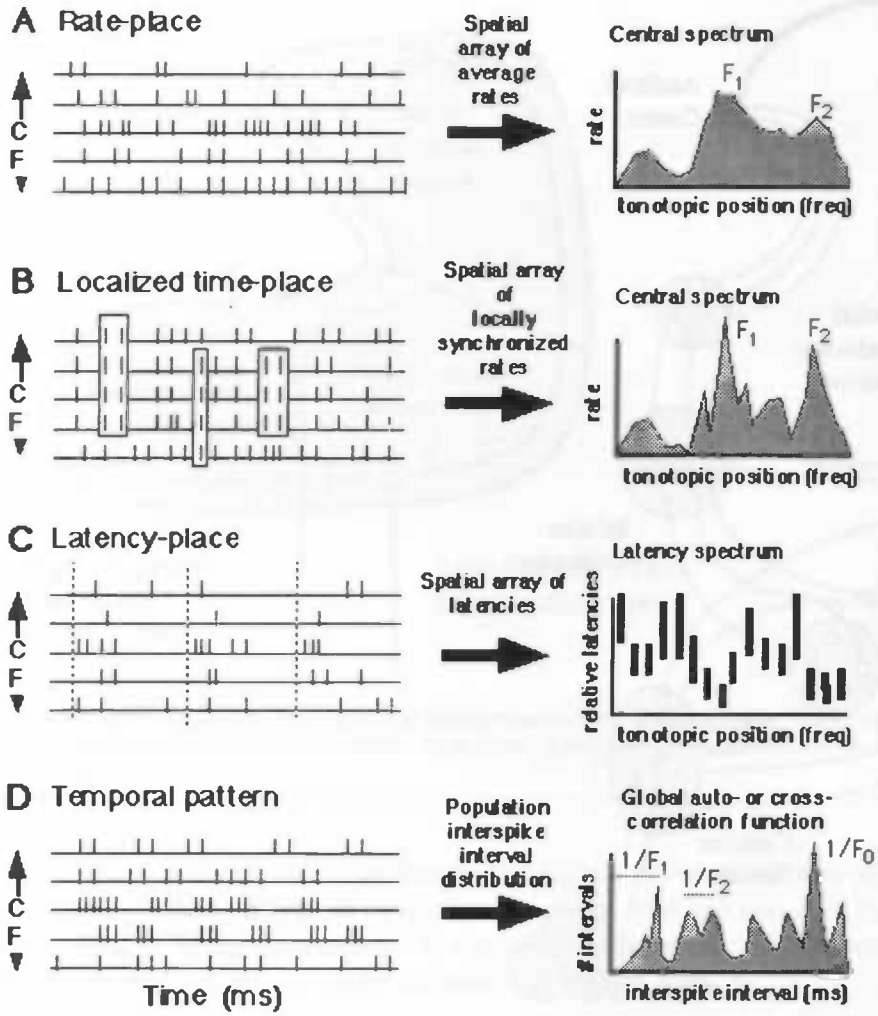


Figure 3.13 Four general neural schemes for encoding acoustic spectra. Left: hypothetical spike trains of the auditory nerve. Right: central auditory representation. (A) Place-based coding scheme using average discharge rates. (B) Place-based coding schemes using synchronization between spatially adjacent channels. (C) Latency-place representation. Vertical bars in th erelative latency map indicate a reange of latences in each frequency channel. Shorter latencies with smaller intensities signal higher intensities. (D) Global temporal pattern coding through population interspike interval statistics. These schemes are not mutually exclusive, and could be potentially combined within neural processing structures.

To summarize, it is likely that the AS employs both representations in concert. Both the average firing rate of individual fibers and their fine-time firing patterns are important for the *complete* representation of speech. The fact that AN fibers are able to lock to stimulus harmonics allows for the representation of formant structure to be passed to higher structures. This representation is more robust (at medium to high stimulus levels and/or in the presence of noise) than is possible on the basis of a decomposition of the stimulus into frequency channels by the local innervation of the BM and the tonotopic organization of the AN. An overview of possible neural schemes for encoding acoustic spectra is depicted in figure 3.13.

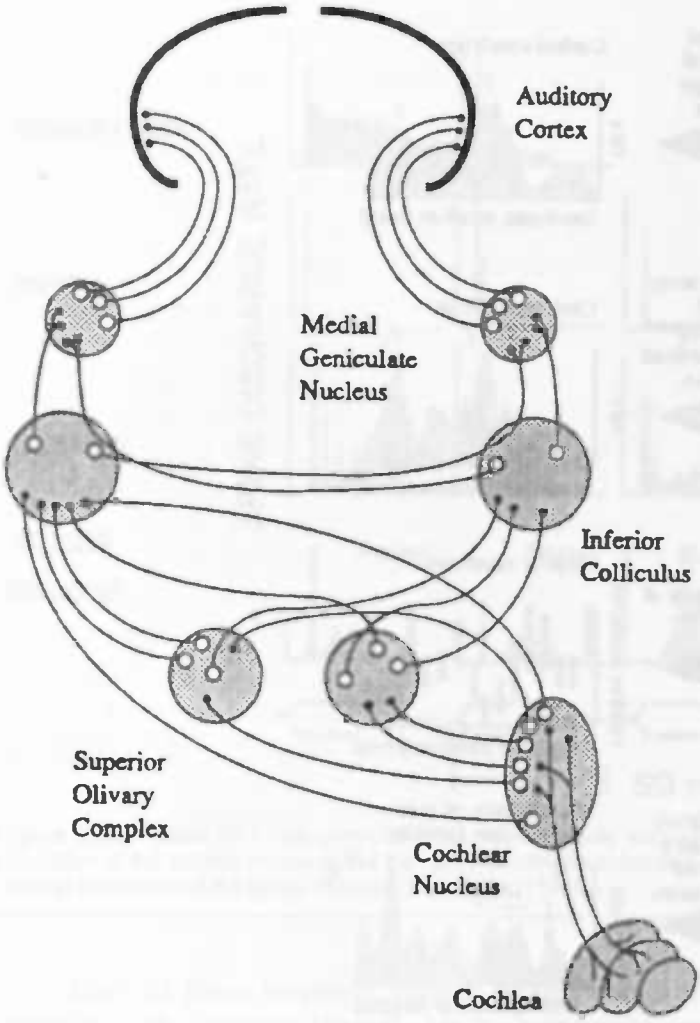


Figure 3.14 Simplified illustration of the ascending auditory pathway depicting the principal nuclei and fiber pathways of the central auditory system. Only pathways from the right ear are shown. Neurons are illustrated with small open circles (o) while synaptic connections are displayed as small filled circles (•). [From Morgan (1991)].

3.3 The cochlear nucleus

The first opportunity for spectral information processing occurs in the *cochlear nucleus* (CN), the obligatory terminal zone of afferent fibers: all fibers of the eight nerve project in an extensive and orderly fashion to the CN, a cluster of neurons residing in the *brainstem* (see figure 3.14). The cochlear nucleus can be divided into a number of distinct subsystems as depicted in figure 3.15:

- the *ventral cochlear nucleus* (VCN), consisting of:
  - the *anteroventral cochlear nucleus* (AVCN)
  - the *posteroventral cochlear nucleus* (PVCN)
- the *dorsal cochlear nucleus* (DCN)

Again, tonotopic organization of the BM is preserved in the termination pattern of the eighth nerve fibers, and each fiber has multiple termination points in each region of the CN. Therefore, each subdivision receives a complete representation of the frequency map preserved in the AN. Fibers at the base of the cochlea (representing high-frequency components) penetrate deep within the nucleus, and fibers from the tip of the cochlea (representing low-frequency components) branch near the surface of the CN.

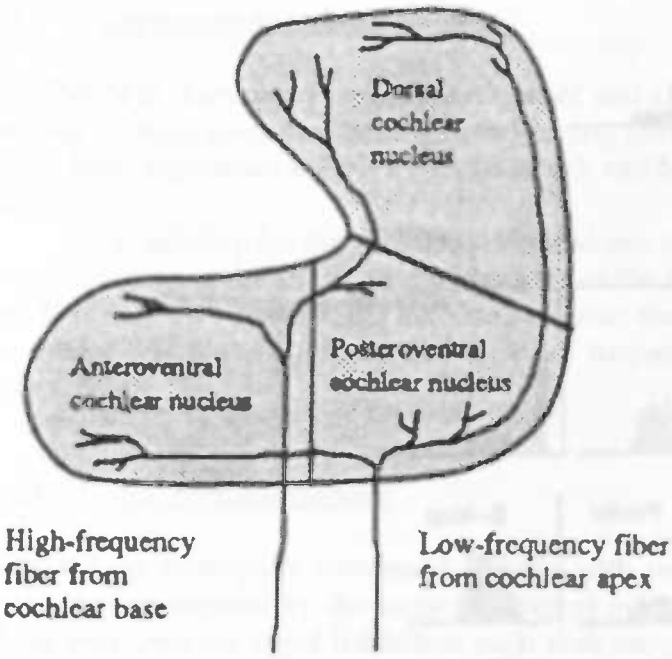


Figure 3.15 The three principal regions of the cochlear nucleus (CN) are shown. Each region receives a geometrically organized arborization of the eighth nerve. Within each region, a laminar organization reflects the tonotopic organization of the basilar membrane.

Furthermore, the branches of the eighth nerve terminate on morphologically distinct cell types with a variety of synaptic structures that are thought to serve as the basis of distinct cellular firing properties. On the basis of characteristic response properties of CN cells the following cell types can be defined (see figure 3.16):

- *Primarylike* cells with response properties (shown in PSTHs) similar to those observed in AN fibers, i.e., showing adaptation effects: high firing rate at stimulus onset, followed by a rapid decline to a constant level for the remainder of the stimulus, sometimes with a brief pause in-between.
- *Onset* cells with a burst of activity at stimulus onset which immediately decays to very low activity for the duration of the stimulus.
- *Chopper* cells showing responses similar to primarylike responses, but with a modulation (or chopping) of the firing rate during the steady-state activity that is synchronized with the stimulus onset. This modulation may decay rapidly during the stimulus, or last for the duration of cell activity, and the rate of chopping may vary from cell to cell and is not related to phase-locking in the afferent auditory nerve fibers.
- *Pauser* cells which have a strong onset spike followed by an extended period of silence lasting at least 5 ms in duration, and is then followed again by a steady-state response.
- *Buildup* cells that are characterized by an absence of any activity at stimulus onset, but a gradually increase to a steady-state.

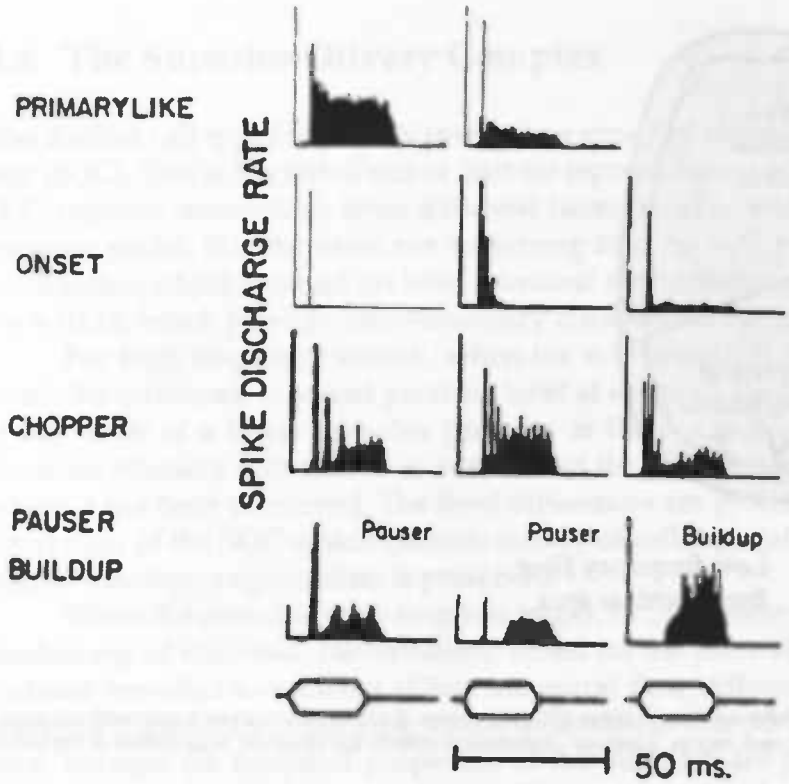


Figure 3.16 Typical PST histograms of single neuron activity within the cochlear nucleus. Each PST was recorded during stimulation at the optimal frequency for the cell. The stimulus consisted of a 25 ms tone repeated at 50 ms intervals as illustrated at the bottom of the figure. [Adapted from Young (1984)].

Each of these response types[34] of cochlear nucleus' cells might be considered an abstraction of the original stimuli. All the response types are found in each of the major subdivisions mentioned earlier, but there is a significant spatial variation in the distribution of response types, probably reflecting differences in the functional role of these regions. In general, there is a gradient of increasing response complexity from the AVCN to the DCN. The exact role of each different response type in the processing of complex stimuli is still a matter of investigation.

3.3.1 Anteroventral cochlear nucleus

Here, the rate-place code as well as the temporal-place code present in the AN is preserved. It seems therefore that the AVCN is specialized for the preservation of spectral and temporal properties of AN input. Activity in the AVCN is dominated by primarylike responses. It has been shown that cells with this response type are able to phase-lock to formant frequencies in vowel stimuli, which is not the case for chopper cells (Sachs, 1988).

[34]Note that the classification of these response types is on the basis of temporal properties they exhibit in response to short tone bursts at best frequency. Because such temporal patterns are heavily complicated by inhibitory processes in the DCN, DCN neurons are sometimes described with an alternative unit classification scheme. Here neurons are described as Type I through Type IV on the basis of frequency- and level-dependent patterns of inhibition. However, regardless of the classification scheme used, in both the VCN and the DCN the defined physiological response types have been correlated with morphological cell types, thus providing a framework for studying the significance of anatomically-distinct entities in auditory information processing (Young & Yu, 2000).

3.3.2 Posteroventral cochlear nucleus

The PVCN is dominated by cells with onset and chopper responses. The posterior portion consists of onset responses having broad tuning curves. It is assumed that the onset response arises from significant inhibitory input which can be either locally generated, or external to the PVCN.

A possibility for the functional significance of the PVCN in speech processing is that it is specialized to detect the sweep direction of the frequency modulation (FM) of the stimulus. This may play a role in the detection of onset times for *vocal inflection*, which is related to the intonation pattern, where the stimulus frequency is changing in a specific direction (Cant, 1984).

3.3.3 Dorsal cochlear nucleus

The DCN is principally composed of cells with pauser and buildup responses. These responses are governed by the large number of interneurons forming complex local circuits. These may provide rapid inhibition such that initial activation is suppressed (Cant, 1984). Also contributing to the complexity of DCN cell responses is the significant innervation from OHCs of the cochlea, and the input that these cells receive from higher centers, including the superior olivary complex and the inferior colliculus.

In contrast to the AN fibers with a very narrow dynamic range and an easy saturation in high-noise conditions, the cells in the DCN are less sensitive to high-noise levels and maintain their sensitivity to tones in high-noise environments. It may be that the local inhibition of the DCN plays a role in this dynamic range preservation through the contribution from inhibitory cells responsive to neighboring frequencies. A possibility regarding the *functional* role of the DCN is that it processes "ongoing activity" in the auditory system.

The differences in organization between the different subdivisions of the cochlear nucleus, suggests that the CN neuron classes play functionally-distinct roles in encoding biologically-relevant spectral information. Research has already indicated that the average discharge rate of an auditory neuron can convey spectral information to other neurons using two different mechanisms. CN neuron types whose discharge rates are well-correlated with spectral level at BF may be involved in a rate-place representation of spectral shape. It is suggested that a tonotopic array of such neurons can send a snapshot of the stimulus spectrum to higher-level auditory centers across a distribution of discharge rates which can be of relevance for the encoding of speech in the auditory nerve (Delgutte & Kiang, 1984) and VCN (Blackburn & Sachs, 1990).

Alternatively, CN neurons may be involved in "feature detection", a mechanism by which the activity across a tonotopic array of neurons is strongly influenced by the presence of wideband spectral features. For example, in the DCN, neurons are found which appear to be robust wideband feature detectors, where a rate-representation of spectral notch location is maintained by enhancing the depth of the notch representation at BF through complex nonlinear inhibitory processes. As a result, the discharge rates of these neurons are prevented to rise far above spontaneous rate at high stimulus rates whereas linearly-behaving (narrow-band) VCN neurons may already become saturated at these higher levels (Yu & Young, 2000).

The fact that different neurons are capable of representing different aspects of the stimulus, is known to be characteristic of sensory systems. This property of multiple representation of the (auditory) environment therefore plays an important role in auditory theories.

3.4 The Superior Olivary Complex

The distinct cell types of the CN project to a group of neurons called the *superior olivary complex* (SOC). This is the initial site of *bilateral* representation of the acoustic environment as the SOC receives connections from *ipsilateral* (same side) as well as *contralateral* (opposite side) cochlear nuclei. It is therefore not surprising that the SOC plays an important role in sound localization which is based on both *interaural time differences* (ITD) and *interaural level differences* (ILD) which provide complementary information for lateralization.

For high frequency stimuli, when the wavelength is shorter than the diameter of the head, the difference in sound pressure level at each ear, the ILD, can be used. This difference is the result of a lower stimulus pressure at the ear that is being shadowed by the head. Stimulus intensity information is available in the SOC as it is encoded in the AN and the CN where it has been preserved. The level differences are processed in the *lateral superior olive*, a subsystem of the SOC which consists mostly of cells specialized for high-frequency stimuli. Again, tonotopic organization is preserved.

When the stimulus wavelength is larger, i.e., for lower frequencies, there is no effective shadowing of the head. Nevertheless, based on the *phase difference* that results because the distance travelled to each ear differ, interaural time differences (ITD) provide the required source of information regarding sound localization. The SOC has access to this phase difference, because the temporal properties of the stimulus are preserved in the AN and in the cells with primarylike responses in the CN, in particular the AVCN. The portion of the SOC to which the cells in the AVCN project is the *medial superior olive*, and through *coincidence detection* in the activity from the AVCN, the ITDs are computed. The use of phase information is not accurate for high frequencies, because of (1) the loss of phase-locking for higher frequencies (above 5 kHz), and (2) the ambiguity in lateralization that results when the stimulus contains components with wavelengths smaller than the ITDs (corresponding to frequencies above approximately 1400 Hz in humans).

The fact that the auditory system uses these complementary sources of information in sound source localization is known as the *duplex theory of sound lateralization*. Again, multiple representations of stimulus information exist. Nevertheless, without the use of intensity information humans are also able to lateralize complex high-frequency stimuli, possibly by focusing on the *modulation envelope* of the stimulus, which is possible through the presence of low-frequency modulation (with a period of 1/F0) of the high frequency components. This illustrates the flexibility with which available information can be applied and the advantage of multiple representations.

3.5 The Inferior Colliculus

Fibers from cells within the cochlear nucleus and the SOC project to the *Inferior Colliculus* (IC). This cell group receives bilateral input from the SOCs, and contralateral input directly from the CN. The convergence of fibers from the CN and the SOCs suggests that the IC serves a special role in localizing sound sources which consist of complex temporal variations (especially because cells in the PVCN are sensitive to the sign of an FM sweep).

Though the fibers of the AN and the cells of the CN are able to represent the temporal variations of stimuli up to several kHz, it has been found that the sensitivity of more central structures of the AS to variations in the temporal domain decreases in progression to the cortex. The manner in which the auditory system represent longer duration temporal structure is of considerable importance in speech processing.

For instance, amplitude modulations of speech signals in the range of 10-20 Hz often occur as a result of phoneme transitions. Syllables produces modulation of speech with a frequency of about 5 Hz, and more complex units of speech (words, sentences) produce even slower amplitude modulations. The IC seems to participate in representing this kind of information. For instance, it has been found that IC cells display modulation frequency selectivity, i.e., they show *best modulation frequency* (BMF) responses over a wide range of modulation frequencies (in the cat over a range of 4 Hz to about 1 kHz, (Schreiner, 1988)).

The neurons of the IC are organized into concentric sheet which form spherical layers, where each layer represents a single optimal frequency of stimulation. Thus, the tonotopic organization characteristic of the preceding nuclei and the auditory nerve is retained. This is combined with a similar map for BMF, and the organization of these two maps is such that iso-frequency bands are orthogonal to one another. This extremely fine regularization of temporal representations has not been observed in other nuclei of the AS, and its functional significance is not known. However, it has been proposed that the IC is specialized for the representation of pitch (Langner, 1981).

3.6 The Medial Geniculate Nucleus

Neurons of the inferior colliculus project to the *medial geniculate nucleus* (MGN), primarily to cells within the ipsilateral MGN, but some cells extend their axons to the contralateral MGN. The MGN resides in the *thalamus*, and it is the last opportunity for filtering the acoustic information before the auditory pathway converges onto the auditory cortex. In fact, the projection area of the MGN *defines* the auditory cortex. There are four subdivisions within the MGN, the ventral nucleus, the posterior group, the combined dorsal and ventral areas, and the medial division.

In the *ventral nucleus* and *posterior group* the spectral features of the stimulus are maintained in detail comparable to that found in earlier representation: cells are tonotopically organized, and are as sharply tuned in the frequency domain as the cells in the IC and the CN. From this section of the MGN, cells project to bands of cells in the auditory cortex which are either excited by input from both ears, or inhibited by input from the ipsilateral and excited by the input from the contralateral ear. Because these projections are segregated from each other, this leads to combined binaural and tonotopic maps.

The combined *dorsal* and *ventral areas* are more broadly tuned. Furthermore, they display a significant delay of several hundreds of milliseconds after stimulus onset in response to simple stimuli. A possibility is that these cells respond only to more complex stimuli than those commonly used (i.e., broadband noise and single frequency tones). Though it appears that the neurons of this divisions are not tonotopically organized, they *do* project to tonotopically organized cells in the auditory cortex.

Finally, the *medial division*, also does not appear to be tonotopically organized, but the ascending inputs to the cortex *are* organized into high- and low-frequency projection systems. The cells in this division have heterogeneous response properties, and display a range of frequency selectivities. In addition, some cells respond quickly after stimulus onset while others show significant delays

Concerning the temporal representations in the MGN, it has been demonstrated that MGN cells can also fire in synchrony with the temporal fine structure of the stimulus. However, progression to these more central structures leads to poorer temporal resolution. For example, in the squirrel monkey analysis of MGN cells indicates best modulation frequencies in the range of 16-32 Hz whereas in the inferior colliculus of the same species it is 32-64 Hz. This suggest that these structures are more specialized in representing more global features that change more slowly.

As has been mentioned, the MGN is the primary thalamic waystation before fibers arrive in the cortex. Not surprisingly, it is difficult to identify simple stimuli at this level of processing. In addition, descending input from the auditory cortex may also play an important role in the responsiveness of cells in the MGN. Nevertheless, it has been shown that some portions of the nucleus seem to retain all of the properties of the prior processing centers, with possibly some loss of temporal resolution. Other regions elaborate on the earlier representations. So, the MGN acts both as a *relay station for prior representations* and as a *generator of new representations*. Off course, as has already been shown, this also holds for other nuclei of the auditory pathway.

It has been proposed that the auditory system employs two *functionally distinct* pathways (Weinberger & Diamond, 1988).

(1) One pathway accurately conveys all the information necessary to *characterize* acoustic events for which the ventral portions of the MGN are the relay station. This pathway "simply" encodes the physical qualities of the stimulus.
(2) Another pathway is characterized by the absence of cells displaying frequency selectivity and the (apparent) absence of tonotopic organization which is the pathway where the cells of the medial division are found.

The cells of this second pathway display *plasticity* while those of the first do not. Therefore, it is argued that the degree of plasticity is inversely proportional to the fidelity of representation of the acoustic stimuli in portions of the auditory system. The second pathway is known to display *associative learning*, and it allows the auditory cortex to selectively label stimuli with perceptual qualities. As this already occurs in the sub-cortical nuclei, these nuclei probably play an essential role in the perception of the acoustic environment, not just in conveying its physical parameters.

These different pathways have also been characterized as respectively the "where" and "what" pathway. In the "where" or "binaural" pathway the information coming in from the two ears is processed for sound localization. In the "what", or "monaural" pathway the spectral and temporal features of the stimulus are processed. This pathway supposedly cares little about the spatial aspects of the stimulus, and mainly focuses on identifying and classifying different types of sounds. However, it should be noted that despite the apparent dichotomy of these two processing pathways, the same types of acoustic cues may be important for the analyses that occur in each. E.g., spectral information is used in the "where" pathway for determining a sound's elevation and temporal information used for pitch perception in the "what" pathway is also used in the "where" pathway for determining a sound's azimuth (horizontal location). The existence of these different pathways reflect the idea that the two main functions of hearing lie in auditory communication and in the localization of sounds, and that information-processing relevant for performing these functions can take place relatively independent.

3.7 The Auditory Cortex

The primary auditory cortex is located in the *superior temporal gyrus*, a portion of the *temporal lobe*. The cells in the auditory cortex are binaurally driven and display the same sensitivity to time and intensity differences as observed in mid-brain nuclei, with cells in each hemisphere apparently being used for sound localisation in the contralateral half of the auditory field. The precise organization of a large cortical area called *A1* is of particular interest.

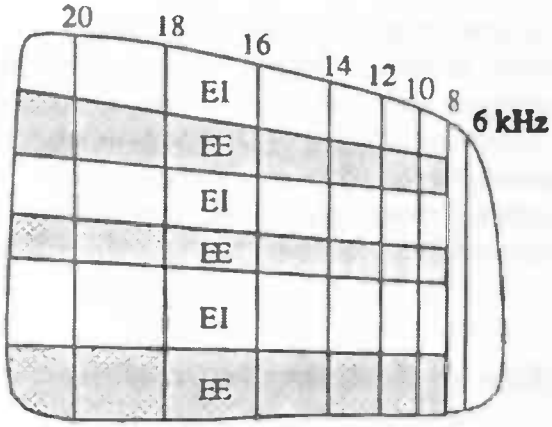


Figure 3.17 An illustration of the auditory region A1 of the cat. Tonotopic bands are shown with binaural bands intersecting at right angles. The EE bands correspond to cells excited by input from both ears. The E1 bands correspond to cells which are excited by the opposite side ear and inhibited by the same side ear. [From Morgan & Scofield, 1991].

Within the cat for example, this area is organized into tonotopic "slabs": cells at various depths, but the same location, respond with similar CFs. These constant frequency slabs are traversed at right angles by binaural slabs of cells. These binaural slabs consist of neurons that are excited by stimuli from both ears (EE band), and neurons that are excited by stimuli from the contralateral ear and inhibited by stimuli from the ipsilateral ear (E1 band). This is illustrated in figure 3.17.

Regarding the response types in the auditory cortex, it is found that many cells of the auditory cortex do not maintain discharges for the duration of a stimulus, i.e., their responses are often *transient* and rapidly *habituate* to a given stimulus. They appear to be optimally sensitive to complex stimuli and often do not respond well to sinusoidal tones. However, many cells are sensitive to frequency modulation (FM) stimuli, e.g., they show sharp tuning for either positive or negative FM sweeps. The cells of the auditory cortex are much more selective to FM stimuli than the cells of the CN which also show FM selectivity.

As has been mentioned earlier, when progressing from the peripheral to central structures of the auditory system, the temporal sensitivity decreases.[35] The temporal resolution of cortical cells is again lower than that of neurons within the IC and the MGN (Schreiner, 1988). Furthermore, these cells appear to be specialized for the spectral components of relatively infrequent stimuli. Investigations of the ability of cortical neurons to encode the short-time structure of amplitude modulated tones, shows that best modulation frequencies (BMFs) lie in the range of 4-16 Hz for the squirrel monkey (Muller-Preuss, 1987), and 10-28 Hz for the cat. There does not seem to exist an organized BMF map in the auditory cortex though. The frequency range of the temporal detection properties of cortical neurons as they are found in cat is centered around the same rate that is characteristic for the repetition rate of speech units (syllables, phonemes, and words), possibly because vocalizations in cat have a similar repetition rate (Schreiner & Langner, 1988).

The role that cortical cells play in the perception of sound and speech remains to be determined. The temporal characteristics of cortical neurons imply that cortical processing is specialized for complex and relatively infrequent spectral events. As cortical neurons are the target of both pathways mentioned earlier, it is likely that they complete the integration of perceptual and physical properties which begins in the MGN.

[35] This decrease of temporal sensitivity is consistent with an increase in the length of the *temporal integration window*.

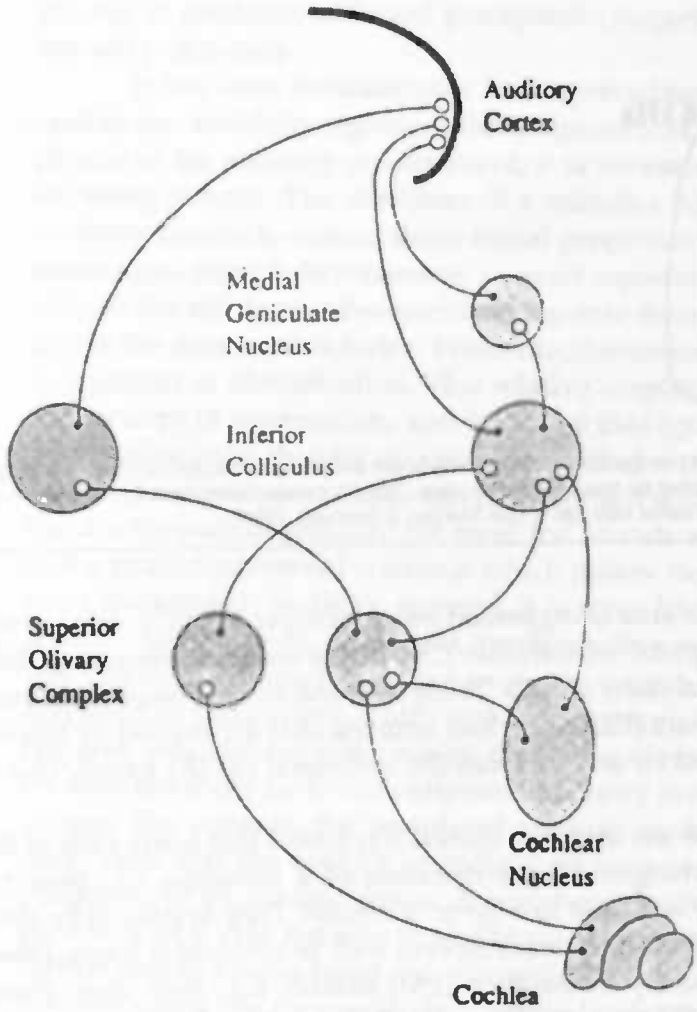


Figure 3.18 Schematic representation of the recurrent pathways in the central auditory pathways. Only the principal pathways and nuclei are shown. Cell bodies are represented with small open circles (o), while synaptic connections are displayed as small filled circles (•). [From Morgan and Scofield, 1991].

3.8 The descending auditory pathway

One interesting fact characterizing the central auditory system is its extensive recurrent connectivity. The descending pathway has the same complexity as the ascending pathway. A highly simplified schematic illustration of a portion of this pathway is depicted in figure 3.18. It can be seen that the following descending paths exist:

- Neurons in the auditory cortex extend their axons back to the cells of the MGN and to cells of the IC on both sides of the brainstem.
- Cells within the IC extend axons to cells of the SOCs, again on both sides of the brainstem, and directly to cells within the CN.
- Fibers emanating from the SOC terminate within the CN, and directly on the OHCs within each cochlea.

The influences that these descending fibers have on the various nuclei has been shown to be both excitatory and inhibitory (Pickles, 1982). For example, when portions of the SOC are stimulated, this leads to an increase in the activity of cells within the cochlear nucleus, whereas stimulation of the lateral portions of the SOC has been found to have an inhibitory effect on the CN.

From a neurobiological perspective, it has been considered very likely that this recurrent pathway plays not only a role in *selective attention*, but also in the *control of perception of stimuli* as a *function of context*. One suggestion is for example, that the recurrent connections from the medial SOC to the cochlear outer hair cells act as a frequency-dependent *gain control*. The recurrent connectivity may therefore be central to maintained perception of complex stimuli in high-noise environments (Ghitza, 1988). Interesting as these ideas might be, at this point, the functional significance of the recurrency in the auditory pathway is not known.

3.9 Auditory neuroethology and speech processing

The overview given so far, is mainly based on an integration of the knowledge obtained within neurobiology and neurophysiology where the emphasis lies on the structure of the auditory system (though attempts are made to relate this to the possible functionality). Another interesting source of knowledge comes from *auditory neuroethology* which is based on the idea that the vocal and auditory systems have evolved together for acoustic communication, under the influence of the acoustic environment. The auditory system has become specialized for receiving and processing the acoustic signals that are the most important for the survival of the species. Therefore, biologically important sounds are used as stimuli to explore auditory mechanisms underlying species-specific behavior, thereby paying more attention to the neural mechanisms that lead to the functional organization of the auditory system.

It is impossible to directly assess the processing that takes place in response to speech stimuli within the human auditory system. Since the representation of speech stimuli by the auditory periphery in animals (such as the macaque monkey, pig or cat) are probably very similar, it is definitely useful to perform animal studies to gain further insight on human auditory information processing. On the other hand, since human speech is unique and highly specialized, the processing of speech sounds by the human auditory center must be unique. As the auditory system must be functionally organized to fulfill the species-specific need, an animal's auditory center must be specialized quite differently from that of humans. This makes studies on the processing of speech sounds by the auditory center in animals much less viable. It is therefore believed that neuroethological studies on different types of animals, including "auditory specialized" animals, will eventually contribute most to the understanding of the basic mechanisms for speech sound processing in humans. Studying the auditory systems of different types of animals contributes by finding the neural mechanisms that are shared among these animals. The reason for using animals that are specialized in their auditory behavior (such as the mustached bat which has a velocity-sensitive echolocation system) is that these auditory systems will show the upper limit of specialization of the neural mechanisms of that behavior.

3.9.1 Information-bearing parameters

Though speech sounds are much more complex than animal sounds, both types of sounds contain similar basic components, that is, *information-bearing* elements. These elements and the relationships among them are characterized by particular values of so-called *information-bearing parameters* (IBPs). The central mechanisms shared by different types of animals are the production of neurons tuned to IBPs, complex-sound processing by combination-sensitive neurons, and anatomical parcellation for the representation of IBPs. These IBPs are related to those aspects of the incoming stimuli that contribute to the ability to discriminate between meaningful stimuli. They are therefore related to *discrimination pressure*, which is determined by the number and similarity of biologically significant sounds that must be discriminated

among. It produces different perceptual categories that may be coded genetically and/or learned postnatally.

It has been demonstrated in the preceding sections of this chapter, that the auditory system can faithfully represent the temporal and spectral signal properties. Given the complexity of the auditory environment, it is necessarily to have an accurate description of the incoming stimuli. The detection of a stimulus belonging to a particular class requires the auditory system to extract those signal properties that allow for the discrimination between meaningful stimuli. In this sense, an exact representation of stimuli is important in so far as it adds to the ability to subsequently categorize them. Thus, for stimulus detection, discrimination is the determinant factor. When discrimination pressure is high, one might also speak of recognition or identification. This relative emphasis on discrimination is directly related to the concept of information, and indicates that (groups of) neurons fire such as to optimize information gain, thereby maximally reducing entropy.

Since complex sounds can be expressed within multidimensional continua, the higher the discrimination pressure the more the sounds will be confined to specific, narrow ranges of the multidimensional continua which makes them more stereotyped and discrete. Within most mammalian auditory systems, it is more likely that distributions of different types of acoustic signals show significant overlap within any single parameter thereby making them less discrete. However, they are probably unique in the particular *combination* of temporal and spectral parameters, which makes them more like (*context-sensitive*) *feature detectors*. This requires very input-specific tuning responses of neurons. For mammals, this kind of selectivity does not show up in the peripheral auditory system, but only until in the central auditory system. For instance, the peripheral auditory system has an anatomical axis for frequency only, such that the activity of individual peripheral neurons cannot uniquely express the properties of an acoustic signal. A peripheral neuron tuned to 2 kHz responds not only to a pure tone of 2 kHz but also to an FM sound sweeping across 2 kHz regardless of sweep direction and to a noise burst containing 2 kHz regardless of bandwidth. The properties of an acoustic signal are appropriately expressed only by the spatiotemporal pattern of the activity of all peripheral neurons. The multiple cochleotopic representation that is prominent in many central auditory systems suggests that separate auditory areas are concerned with the representation of different types of auditory information. The question therefore is what kind of information is represented in each area, and how is each area functionally organized? There are a number of (complementary) hypotheses regarding the functional organization of the auditory center (see figure 3.19) where the validity of each hypothesis depends upon the types of auditory information and species (Suga, 1982):

- *Amplitude spectrum hypothesis*: An acoustic signal is represented by the spatiotemporal pattern of activity of "nonspecialized" neurons arranged along the coordinates of frequency versus amplitude. This typically occurs at the periphery, where the frequency of an acoustic signal is expressed by the location of activated neurons and its amplitude by their discharge rates. This hypothesis is considered to be too simple to explain neural mechanisms for speech-sound processing.
- *Detector hypothesis*: A biologically important acoustic signal is represented by the excitation of detector neurons that selectively respond to that particular signal. Off course, different types of detectors are arranged in a particular spatial pattern. Essentially, it is assumed that there exists a one-to-one correspondence between signal and excitation of a detector neuron or neurons in a single cortical column. However, the neurons found thus far are not specialized enough to be called detectors.

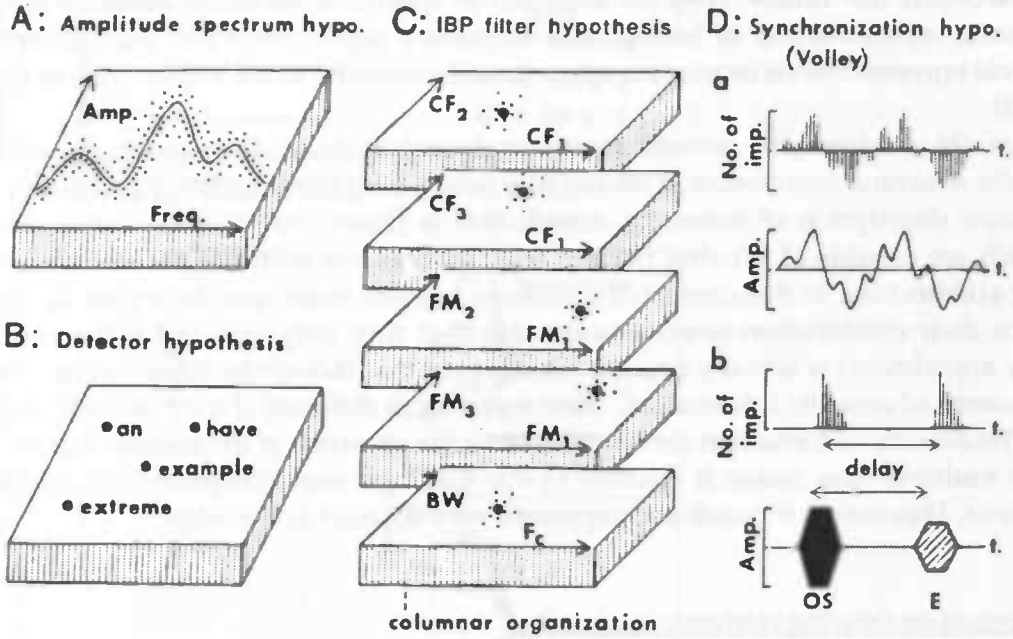


Figure 3.19 Four working hypotheses for the representation of auditory information by nearal activity in a hypothetical center. (A) The amplitude spectrum hypothesis. (B) The detector hypothesis. (C) The IBP filter hypothesis. Coordinates are formed by neurons tuned ot information-bearing parameters (IBPs) that characterize biologically important signal elements. F_c and BW are the center frequency and bandwidth of noise bursts, respectively. In the hypotheses A-C, the properties of acoustic signals are represented by the spatiotemporal pattern of neural activity. However, the response properties of individual neurons and the interpretation of their functions are different according to the hypotheses. (D) The synchronization hypothesis. In D_a, the lower trace represents a sound wave; the upper trace is the compound period histogram of a single neuron response to it. In D_b, the lower trace represents an orientation sound (OS) and an echo (E); the upper trace is the PST histogram of a single neuron response. [From Suga, 1988].

- *Information-bearing parameter filter hypothesis*: All neurons in the auditory system, including the peripheral ones, act as filters: they *correlate* acoustic signals with their filter properties, that is the information that is stored within them. The degree of correlation is expressed by the magnitude of the output of the filters, thus the neurons are maximally excited only when the properties of acoustic signals perfectly match their filter properties, i.e., stored information. Within the IBP filter hypothesis, specialized neurons that express the outputs of neural circuits tuned to particular (combinations of) IBPs are represented in the auditory center. IBP filters correspond to spatiotemporal patterns of activities of these specialized neurons that acts as a cross-correlator. Different types of IBP filters are aggregated separately in identifiable areas of the auditory center. This hypothesis falls between the two earlier hypotheses.
- *Synchronization hypothesis*: When information-bearing elements are lower than 5 kHz (the upper-limit of neural phase-locking), peripheral neurons produce discharges synchronized with the sound waves. The envelope of a compound period histogram of a neural response thus reproduces the stimulus waveform. According to this hypothesis, neurons in the auditory center represent acoustic signals by synchronous discharges. For instance, pitch can be expressed by such synchronized activity.

The functional organization of the auditory cortex does not necessarily follow a similarity in gross anatomy and phylogeny. Rather, it follows the properties of biologically important sounds. The IBP filter hypothesis is therefore the most likely candidate for the representation of speech stimuli in the auditory cortex, and has been supported by data obtained from different types of animals, including auditory-specialized animals. For instance, in the auditory cortex of the mustached bat, different kind of IBPs characterizing complex acoustic signals are separately and systematically represented by combination-sensitive neurons that are

tuned to particular IBP values. They are arranged in specific areas of the auditory cortex for the systematic representation of biologically important signal variation. Furthermore, the size of neural representations or maps is apportioned according to the importance of the IBPs (Suga, 1988).

Thus, *the structure of the acoustic environment is reflected in auditory processing and is represented in the structural organization of the auditory cortex*. Peripheral auditory processing leads to an accurate description of incoming stimuli that is passed on to more central auditory center which are capable of filtering out the relevant features within a parallel hierarchical processing architecture. In this context, the features become more specific higher in the hierarchy, since their combination-sensitivity ensures that they only respond if the context in which they are relevant is actually present. This means that though features can be based on the same source of acoustic information, their meaning is different if they occur in different contexts. The acoustic information serves as a *cue* to the presence of the feature, but the combination of multiple cues makes it a *feature* (a cue that has been interpreted) of a particular acoustic event. This makes it possible to represent very abstract knowledge.

3.9.2 Information-bearing elements in speech

Communication sounds are commonly characterized by many different parameters: frequency, FM rate, FM depth, amplitude, AM rate, AM depth, harmonics, duration, interval, and so on. Some parameters are IBPs. Though parameters are a continuum and can have any value, only a limited part of the continuum, the information-bearing parameter, is important for each species. For speech-sound processing it is suggested that there are essentially three information-baring elements: formants, formant transitions, and noise bursts or fills (see figure 3.20). Auditory information is carried not only by the acoustic parameters characterizing these elements, but also by other parameters that represent relationships among elements in the frequency, amplitude and time domain. For instance, voice onset time (VOT) reflects a time interval between two acoustic events, and is an important cue for speech recognition (the phonetic boundary of VOT is 22 msec for /b/ vs. /p/, 35 msec for /d/ vs. /t/, and 41 msec for /g/ vs. /k/). Also, information for sound localization is carried not by parameters characterizing a sound, but by interaural time and amplitude differences (ITDs and ILDs). Though nothing is known about the functional organization of the human auditory system, the following speculations based on the above considerations are at least worth considering (from Suga, 1988).

3.9.2.1 Vowel recognition

Vowels might be recognized by combinations of F1, F2, and F3, most likely the combination of F1/F2 and F1/F3. Such frequency-by-frequency coordinates have been demonstrated in the auditory system of the mustached bat. In terms of complex-sound processing by combination-sensitive neurons, it is not at all important whether these CF signals are harmonically related or not, because the question is whether the auditory system contains neurons examining or tuned to particular combinations of two activated locations along a tonotopic axis. At the periphery, the tuning curves for frequency are broader at higher stimulus levels such as the intensity levels at which speech typically is produced, especially for vowels. Therefore, saturation occurs so that the ambiguity in coding of stimulus frequency is very large. In the central auditory system, this ambiguity is reduced in some neurons by lateral inhibition (having its influence primarily at the skirts of the curves), so that these neurons have relatively sharp level-tolerant tuning.

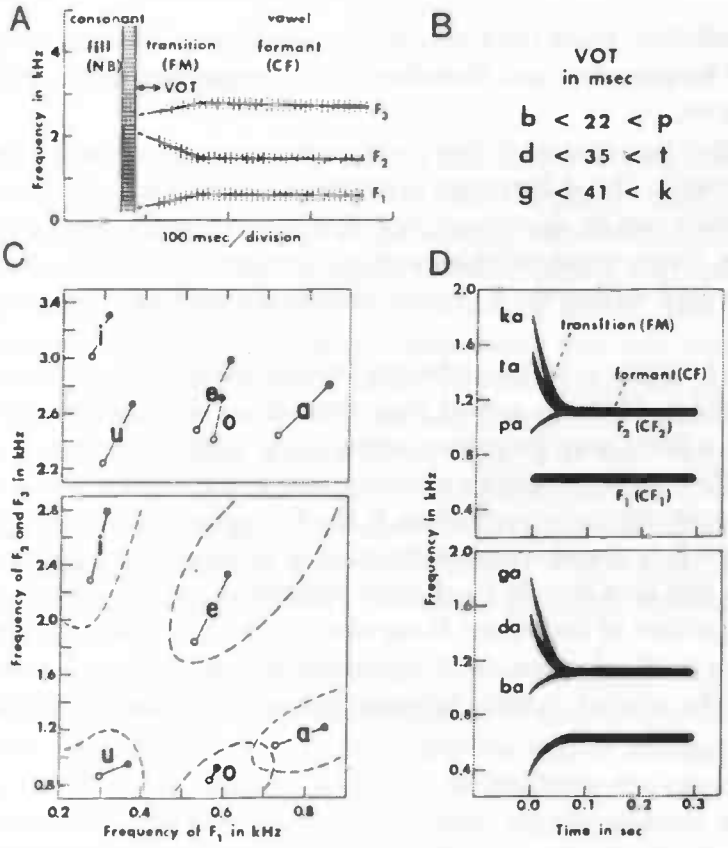


Figure 3.20 Information-bearing elements in human speech sounds. (A) Schematized sonogram of monosyllabic sound shows four types of information-bearing elements: fill or noise burst (NB) component; transition or frequency-modulated component (FM); formant or constant frequency (CF) component; and voice onset time (VOT) or time interval between two acoustic events. F1, F2, and F3 are the first, second and third formants of a vowel, respectively. (B) The phonetic boundary of VOT is given for several phonemic contrasts. (C) The relationship among the frequencies of the F1, F2, and F3 of five vowels (/a, e, i, o, u/). The average frequencies of the formants differ among male (o) and female (•) speakers.The areas surrounded by dashed lines in the lower figure represent formant frequencies for unanimous classification of vowels by listeners. (D) Schematized sonograms of monosyllabic sounds. /a/ changes into /pa/, /ta/, or /ka/ through the addition of different FM components to F2, which in turn become /ba/, /da/, /ga/ respectively through the addition of an FM component to F1. [From Suga, 1988].

The extent of sharpening can differ according to the importance of the signal elements. However, with very high intensities, the problem of saturation remains significant. On the other hand, a distribution of strength of phase-locked responses over the array of neurons appropriately expresses the formants regardless of stimulus levels. Therefore, the phase-lock code incorporated with the place code is far superior to the rate code incorporated with the place code. Since the phase-locking capability of neurons in the primary auditory cortex is limited, the phase-lock code must be translated into a place code at a subcortical auditory nucleus.

A possible mechanism that is capable of such a transformation is a coincidence-sensitive neuron that responds to the impulses from phase-locked neurons, and acts as an AND gate through two axonal branches, one acting as a delay line. Each delay line is adjusted to evoke the delay of an impulse which is the same as the period of the BF of each phase-locked neuron. Thus, a phase-locked neuron tuned to a 1-kHz sound would have a 1-ms delay line, and the coincidence-sensitive neuron is only excited when impulses arrive simultaneously at the two branches. There is an array of coincidence-sensitive neurons associated with different delay lines. Such a neural mechanism allows the appropriate encoding of formants even at high pressure levels. The frequency tuning curves of these neurons are not sharp but may be sharpened by lateral inhibition. For further processing of vowels, the relationships among formants are examined by neurons sensitive to the combination of

F1/F2 and F1/F3. Due to lateral inhibition, these neurons can be excited only by sounds with sharp spectral peaks at two specific frequencies, and therefore not by noise bursts since these also excite the neighboring frequencies.

Formant frequencies and other parameters differ among speakers and are important in identifying both speakers and vowels. The differences in formant frequency is therefore a biologically important variation which can be systematically represented by the frequency-versus-frequency (F/F) coordinates. Since some combination occur more often than others, these combinations are over-represented within the F/F area leading to nonlinear frequency axes.

It has been found that the frequency ratios between the formants of a vowel are nearly constant across speakers (Miller, 1984). A vowel may therefore eventually be processed by neurons that are tuned to a particular frequency ratio. Since such neurons are not suitable for speaker identification, the auditory system must contain two groups of neurons suited for either speaker or vowel recognition. Nevertheless, if the F/F area is assumed to be a neuronal tissue, the activity of which is somehow directly related to vowel recognition, a spatial pattern of neural activity in this area would be directly related to the recognition of both vowels and speakers: the recognition of the vowel is represented by *any* spatial pattern of neural activity occurring within a particular area, and a difference in a particular vowel between speaker is represented by the spatial pattern between neural activities that occurs within this area.

3.9.2.2 *Noise bursts*

Likewise, neurons exist that do not respond to pure tones but to noise bursts and/or clicks, so-called NB-sensitive neurons. A cluster of NB-sensitive neurons with broad tuning curves has been found in the mustached bat. A possible neural mechanism may be facilitation due to simultaneous excitation of presynaptic neurons with different BFs by a noise burst. It may be hypothesized that the auditory system contains a cluster of NB-sensitive neurons to process fills and that the NB area is organized in coordinates representing center frequencies-versus-bandwidths of noise bursts.

3.9.2.3 *Phoneme combinations*

When two phonemes are combined to form a monosyllabic component, formant transitions appear. These transitions reflect frequency modulation (FM) components. Combinations of formant transitions are very important for speech recognition. For example, when a transition is added to the F2 of /a/, dependent on the transition properties, the sound is perceived as /pa/, /ta/, or /ka/. When a second transition is added to the F1 of /a/, it is recognized as /ba/, /da/, or /ga/.

Though FM-sensitive neurons and neurons sensitive to FM combinations have been found in, for instance, the auditory system of the mustached bat, there are actually no neurophysiological data that support a speculation for an area that is specialized in the representation of combinations of transitions. Also, what makes speech so unique in terms of acoustic pattern, is its enormous number of different combinations of phonemes, and therefore also of combinations of transitions. If an area devoted consisting of many arrays of transition-sensitive neurons exists, its functional organization must be unique and complex. However, given the lack of neurophysiological data, speculations regarding this issue will not be made at this point.

3.10 Summary

From the overview presented in this chapter, it becomes clear that the auditory system has access to spectral as well as fine-time temporal characteristics of the speech signal, and that even the spectral *envelope* (representing formants) is adequately represented by the auditory system. It seems that the signal is decomposed, and specialized areas in the auditory system process specific aspects of the incoming signal. This leads to multiple representations of transformed (speech) signal components that can complement each other in representing information present in the signal in order to initiate behaviorally adequate responses. However, because of the (bi-directional) interconnections between different brain regions, these specialized regions should be considered as *semi*-independent. Especially in lower centers, there is a large interchange of information between the two sides of the auditory nervous system. For instance, information for sound localization can be analyzed in early parts where temporal information is more accurately determined. The increasing complexity of the cell responses in higher centers can be traced to the *combination* of the various separated components.

As has been mentioned, auditory research is often conducted on nonhuman mammals by performing measurements with simple stimuli. For the peripheral auditory system, this approach can be validated, but for higher auditory center, this is questionable. Within neuroethology, the emphasis lies on studying the neural mechanisms underlying the processing, representation, and functional organization of species-specific behaviorally-relevant information. This provides a useful source of information, but more importantly, emphasizes the importance of studying behavior in the context of the requirements imposed by the environment. Based on neuroethological studies in auditory specialized animals, it was demonstrated how *the auditory system accurately represents the statistical properties of the acoustic environment*, and that *this is directly reflected in how incoming stimuli are evaluated*. The concept of "meaningful" information, more specifically information-bearing parameters, which is related to discrimination pressure, plays an important role in the understanding of auditory (more generally, perceptual) processing at higher auditory center. Information is not only species-specific, but because of its dependence on previous exposure to (acoustic) stimuli, also carries a subjective component between members of the same species. It seems that certain neurons in the auditory center are optimized such as to maximally reduce the entropy and therefore to increase the information gain. At the same time, however, the auditory system as a whole is also extremely redundant in that it contains multiple representations of the same stretch of acoustic input, and that it can combine these multiple sources of information to estimate a certain signal (or source) property.

Chapter 4 *The Auditory Scene*

This chapter will deal with the problem known as *Auditory Scene Analysis* (ASA). The questions that are addressed are related to the complexity of the real auditory scene in which multiple sound sources are present, and where we are usually interested in the signal stemming from a single source. This aspect of auditory information processing has been extensively studied by Bregman (1990) in his book *"Auditory Scene Analysis, The Perceptual Organization of Sound"* in which the first systematic investigation of the problems that have to be solved in the real auditory scene are described, and possible ways as to how the auditory system (AS) may solve these problems are proposed.

It should be noted though that auditory scene analysis refers to a *function* the AS has to perform, *not* a *mechanism*. Therefore, the obtained experimental findings are often explained in rather *abstract, functional terms*. The physiological mechanisms underlying these functions are usually considered to be research topics for other disciplines. This is not that surprising, since our understanding of information-processing, especially in higher centers of the brain, is still incomplete. This chapter will therefore focus mainly on knowledge derived from the field of psychoacoustics, in which the relationship between the lowest level (the stimulus) and the highest possible level (the sensation, or perception[36]) is described. Nevertheless, some attempts to link the observed findings with the physiological mecha-nisms that might be responsible for them will be made.

Another point to mention is that, to reveal the underlying functionality, relatively simple stimuli under quite unnatural, controlled conditions are used. Given the nature of the research questions that are addressed, this at least casts doubts on the validity of the conclusions that can be drawn from these experiments. Off course, this is a "problem" for most psychological inquiry. Nevertheless, the experiments conducted by Bregman and other scientists working in the field of psychoacoustics do provide many insights regarding human auditory information processing, since they provide a useful *task description* of the system. For a large part, this determines the functional *architecture* of any system (including ASR systems) that has to accomplish the task of auditory scene analysis. These system requirements can in turn be combined with current knowledge about the anatomy and physiology of the mammalian auditory system, with models of human information processing, insights from processes that are at work in learning, memory and perception, etc. (see also chapter 6).

Before addressing auditory scene analysis, some general aspects of perception are considered, and some of the approaches and methodologies for studying perception are introduced in section 4.1. After this, the basic ASA theory as proposed by Bregman will be outlined in section 4.2. This theory, in which a distinction is made between primitive and schema-based grouping of signal components, will be critically reviewed in the light of some illustrative perceptual phenomena related to ASA.

[36] Historically, a distinction is made between sensations and perceptions, where the crucial difference would lie in that perception always refer to external objects, whereas sensations refer to experiences within a person that are not linked to external objects. In practice, however, the term perception is usually used to refer to all experiences caused by stimulation of the senses, and the purpose is to understand how stimulation of the senses leads to the experiences that are part of our awareness of the world around us.

The original ASA theory is primarily based upon the formation of so-called *auditory images* by using repeating cycles of sequences of discrete steady-state tones, frequency glides, and other relatively simple, unnatural stimuli. Therefore, to test the validity of this theory for speech perception, additional experimental findings, in which synthetic stimuli that are abstractions of speech signals are used, will be presented from section 4.9. Of course, using these kind of stimuli within an experimental context illustrates the other side of the caricature that is made of normal speech processing in the presence of multiple sound sources. Nevertheless, it becomes clear that additional principles are needed to explain these experimental data. In the chapter summary, the conclusions that can be drawn from both lines of research will be summarized into a more integrated view of ASA.

4.1 Perception

Many aspects of our mental capabilities like thinking, learning, memory and emotions depend on receiving information about the outside world from our senses. However, one can not study *any* mental process in isolation from the other. Human perception is intricately related to the process of human attention and memory. We will probably not understand one of these until we understand the others. But attention and memory are also related to other systems, to language, to development, to human activity (Lindsay and Norman, p.3).

Nevertheless, the importance of perception in our interaction with the environment is clear. Based on experience with the environment, which we deal with in an *active* manner, in combination with build-in, innate mechanisms, it is possible to build mental representations of the external world and therefore to perceive the world around us.

4.1.1 The importance of the perception of events

Of particular importance in dealing with the environmental scene is the perception of *events*. An event can be defined as *an object undergoing change, such as movement*, or as *changes in structure over time*. Such transformations appear to yield more efficient, stable and veridical perception (Goldstein, 1989, p.301). Regarding auditory perception, it should be realized that we hear *because of* changes in air pressure (air vibrations), which are usually created by the movement of objects, in the case of speech: movements of articulators in the vocal tract.

Events signal changes in the environment for which an appropriate response may be required. The information that is provided by signaling events is therefore intricately related to survival. It may therefore be expected that, through evolution, our perceptual system has been optimized to process the sensory evidence in such a way that behavioral benefit is optimized in combination with the requirements of the environment. This means that the perceptual system deals effectively and adequately with *behaviorally significant* information, present in the environment or abstracted from it, which refers also to the importance of *information-bearing elements* (chapter 3, section 3.9). Important properties of the perceptual system can be derived from the requirements that are imposed by the environment, and that reflect the importance of events. *All sensory systems seem to require changes in stimulation in order to maintain perception.* Steady-state stimuli do not provide information, whereas changes do provide information. Therefore, the *dynamic* properties of environmental stimuli – the *event features* - are likely to provide the most useful sources of information, especially for acoustic signals, which are characterized by their time-evolving patterns.

The most reliable, and therefore probably also the most important, event features are those that can be identified under arbitrary circumstances, in particular in noisy environments. Feature identification often requires an *interpretation* of acoustic cues to the presence of these features. These features are *abstractions* (generalizations) of the details present in the physical stimulus. Because of inherent, but perceptually irrelevant, variations in the physical realizations, these abstract representations allow the system to be more noise robust.[37]

The noise robustness requirement also appeals to the importance of the *redundancy* of the representation of information. One feature is usually signaled by multiple cues. For the interpretation of a cue as signaling the presence of a particular feature, context is always required. Therefore, *nothing is dealt with in isolation.* The integration of the multiple signal

[37] See also chapter 2 concerning the desirability of referring to articulatory simplicity in understanding the process of speech production in relation to the diversity of acoustical cues, and chapter 3 regarding the many transformations on the signal to extract information, and the responsiveness of the auditory system to abstract features of the signal such as formants.

representations that exist while processing auditory information, possibly through the activation of context-sensitive *feature* detectors, allows for the formation of a consistent overall interpretation of the environmental scene (which is usually considered to be reflected in a certain *pattern* of activity within the brain). This way, the system can function properly under many different circumstances despite a (momentary) inability to estimate a certain cue or feature.

Under normal circumstances, all contextual information fits together. Illusions are often the result of contextual cues providing contradictory information to the overall interpretation. This mostly occurs under unnatural conditions, e.g., by manipulating such cues in an experimental context. Nevertheless, experimental findings obtained by such manipulations can provide useful additional insights regarding the mechanisms involved in normal perception.

4.1.2 Research approaches to perception

There are several *approaches* to the study of perception which will be briefly addressed. In general, perception is not considered to consist of a *sequence* of steps, but as a *continuing process* in which different processes are continuously occurring and, in some cases, ((in)directly) affecting one another.

Cognitive processes such as thinking and memory play a role in that they are both an outcome as well as a determinant of the perceptual process. Therefore, in a *cognitive approach* to perception, the focus is on how perception is affected by the meaning of a stimulus and by the subjects expectations (or mental state).

The process of perception involves an interaction between the *data* stimulating the sensory receptors and *information* from past experiences. The latter are the result of *learning* as a result of previous *interactions with the environment*. The *constructive approach* interprets this fact by viewing perceptions as constructed by the observer from perceptual data obtained during *active* observation of the stimulus (i.e., *data-driven*), where the *interpretation* of this data is aided by the observer's *knowledge* of the environment and past *experiences* in perceiving. This is often referred to as *prediction-* or *hypothesis-driven* perceptual processing. The sensory stimulation can be thought of as providing data for hypotheses concerning the state of the external world, mediated by the active role of the observer (for example in vision, by making a series of eye movements in order to fixate on different parts of an object or scene). The acceptation of an activated hypothesis then occurs on the basis of some sort of *likelihood principle*: the object or event that is most likely to be caused by the sensory stimulation is perceived.

The *ecological approach* emphasizes the *active* observer that is moving through the environment. Therefore, perception should be studied in *natural* settings and the stimulation an observer encounters while moving through the environment should be considered. The information present in the environment is enough and mental "calculations" are therefore unnecessary.

Structuralism is an approach that dominated the early years of the 20th century, and that considered perception as made up of many sensations. Though this idea is not accepted anymore, it did give rise to the *Gestalt approach*. Gestalt psychology is characterized by the idea that *"the whole is different from the sum of its parts"*, and attention should be paid to the *overall stimulus pattern*. So, the *global context* plays an important role. In this view, the grouping of parts into larger units (*perceptual organization*) is essential for making sense of the world. Properties of the stimulus determine *perceptual grouping* by the presence of *organizational laws*. These laws should be interpreted as *forces of integration* leading to *simultaneously acting forces of attraction that compete and cooperate* within a *perceptual field*.

Examples of such laws or principles are that of:

- *pragnanz/good figure/simplicity*: every stimulus pattern is seen in such a way that the resulting structure is as simple as possible.
- *similarity*: similar things tend to be grouped together.
- *good continuation* or *completion*: points that, when connected, result in straight or smoothly curving lines are seen as belonging together, and lines tend to be seen in such a way as to follow the smoothest path
- *proximity* or *nearness*: things that are near to each other appear to be grouped together
- *common fate*: things that are moving in the same direction appear to be grouped together
- *belongingness*: one piece of sensory evidence can only be attributed to one perceptual organization, also known as the principle of *exclusive allocation*; a weaker version of this principle states that if a property of an experience is perceived, it is always experienced as a *property of something* in that experience
- *meaningfulness* or *familiarity*: things are more likely to form groups if the groups appear familiar or meaningful

These principles also play an important role in the separation of "figure" and background (*figure-ground segregation*). The Gestalt psychology has played a particular prominent role in the field of visual perception (as can be seen in some of the formulations above), but the above mentioned principles can easily be translated into dimensions suitable for auditory perception.

The next step is to take these principles beyond the descriptional level and focus on the *process* of perceptual organization to make *predictions* about the perceptual outcome possible. It then becomes important to formulate what is implied by, for example, the law of simplicity. What makes one organization simpler than another one? Or, what are the dimensions, i.e., stimulus properties, on which to apply the similarity principle? How is our ability to extract information affected by grouping? What happens when there is an ambiguity in the organization as a result of competition between different principles or signal properties? Modern extensions of Gestalt psychology focus on these kind of questions. Unfortunately, such explanations remain in terms of *functional* descriptions, and usually do not refer to the underlying physiological mechanisms that are responsible for performing these functions. This brings us to a final approach to perception that will be discussed here.

Within the field of *psychophysics* an attempt is made to determine the connections that exist between the physical stimulus and perception. Therefore, the first step is to study the relationship between the physical stimulus and the observer's perceptual response to that stimulus, i.e., the relationship between the beginning and end of the perception process. Often, such descriptions serve as the starting point for physiological investigations. Relations between physiology and perception (*psychophysiology*) can be derived by looking for the relationship between physiological and behavioral data from the same organism. Most psychophysical research focuses on the description of the relation between stimulus and perception.[38] A number of psychophysical methods or approaches will be introduced here, since they constitute the methodology used in experiments on which the ASA theory is primarily founded.

[38] Some researchers seem to be doing little more than quantifying these stimulus-response relations in that they do not seem to make any serious attempts in extrapolating their findings into a broader context to gain further insights in and a deeper understanding of the "true" importance of the phenomenon they are studying.

- The *phenomenological method*: The observer is asked to describe what he or she perceives. This is important, since the first step in any area of perception involves a description of the phenomenon to be studied.
- *Identification* of *information sources in the environment*: The question of what information in the environment contributes to our ability to perceive perceptual qualities is addressed, so this is also based on observation and description.
- *Detection*: Determining the smallest amount of energy needed to detect a stimulus (*absolute thresholds*), thereby defining the *sensitivity* of our senses (sensitivity = 1/threshold), which can be presented in sensitivity or *threshold curves*. An example is the detection of pure tones as a function of frequency and/or stimulus duration (temporal integration). This can be depicted in a frequency-threshold curve, called an *audiogram*. Another measurement, that of *relative thresholds*, provides information on how two or more thresholds compare to one another. For example, in *masking* experiments, the amount of increase of the detection threshold for a certain stimulus when presented in isolation, is measured when the same signal is presented in the presence of another stimulus (see also section 4.5).
- *Matching*: Two stimuli are to be adjusted by subjects so that they appear identical. For instance, by asking a subject to make a judgment of equal loudness[39] with respect to two tones of different frequency, leading to *equal-loudness contours* (often the Loudness Level estimation is made with reference to the intensity of a 1000 Hz tone that has to be adjusted in order to sound equally loud).
- *Scaling*: The questions addressed try to relate the *magnitude of experience* to the *physical intensity* of the stimulus, e.g., by adjusting the level of a 1000 Hz tone until it sounds half as loud as some standard that is used as a reference (like the loudness of a 1000 Hz tone at 40 dB, which is defined as one *sone*).
- *Changing the environment*: It is a property of all sensory systems that exposure to a stimulus of sufficient duration and intensity produces changes in the responsiveness of the system. In *selective adaptation* experiments, a stimulus is presented for a short period of time, or repeatedly presented, and the effect on detection or identification is measured.
- *Identification*: In identification experiments, it is measured how a subject's past history and experience with the world influences its *interpretation* of the sensory stimulus.

These methods constitute the basic methodology of the psychophysical approach to perception. *Psychoacoustics*, a subfield of psychophysics, deals exclusively with the relationship between acoustic stimuli and responses to these stimuli.

It should be stressed that these different approaches are truly *approaches* for studying perception. Findings obtained and conclusions derived within each of these approaches do not have to exclude each other. In fact, as is often the case when combining evidence from different research approaches, these sources of information may complement each other. Sometimes, different approaches can just be interpreted as descriptions of the same phenomenon (or system) on different levels of aggregation, where the eventual task is to formalize the relations between these different levels. For example, by explaining how the higher-level, *macroscopic* properties can be seen as the *implicit consequences* of processes operating at a lower, *microscopic* level. Note that this does not imply a *reduction* of these higher level descriptions to these lower level descriptions (see also chapter 1, section 1.2).

Having introduced some important concepts regarding the study of perception, and the approaches taken to gain a deeper understanding of the perceptual system, we will now turn to the problem known as Auditory Scene Analysis.

[39] Loudness is the perceptual analogue of intensity.

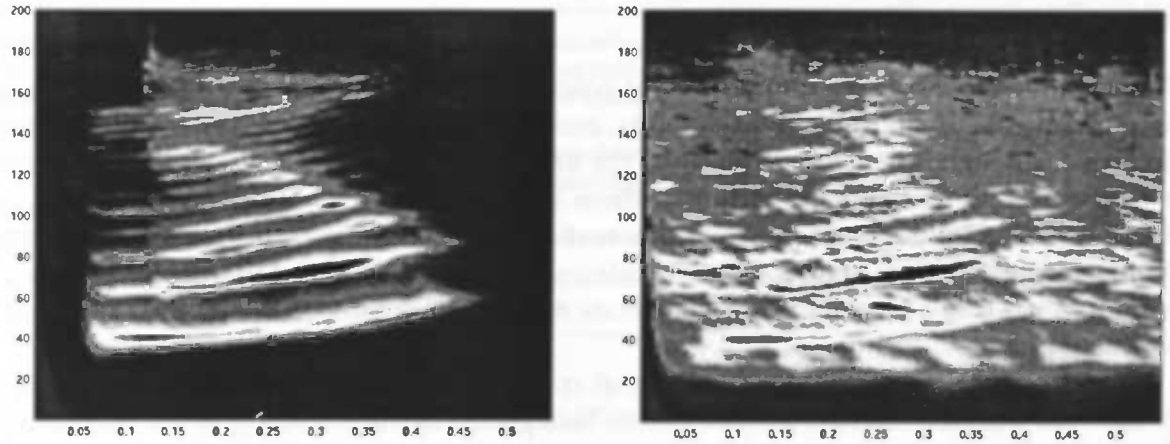


Figure 4.1 The problem of auditory scene analysis. (a) Cochleagram representation of the Dutch word *nul* spoken in isolation. (b) A cochleagram of a mixture of sounds containing the word *nul*.

4.2 Auditory Scene Analysis

Inspired by ideas from Gestalt psychology, and attempting to integrate insights from the several approaches to perception, Bregman (1990), describes the task of perception as *"the process of using the information provided by our senses to derive a useful mental representation of reality from it"*. To do this, it is important that the system is able to "decide" which parts of the sensory stimulation stem from the same environmental event. According to this view, our ability to *recognize* what is going on *depends* on putting the right combination of sensory evidence together.

When a sound source produces a signal, it is rarely produced without the presence of other sound-producing sources, which means that the sensory evidence reaching our ears, and therefore also the signal representation at the cochlea, is more often than not a mixture of signals (illustrated in figure 4.1). The problem of auditory scene analysis can be described as follows:

Although we need to build separate mental descriptions of the different sound-producing events in our environments, the pattern of acoustic energy that is received by our ears is a mixture of the effects of the different events. How can we arrive at the correct combination of information, stemming from one particular source, to allow the reconstruction of a mental representation of one particular auditory event? Or alternatively, which parts of the sensory stimulation are telling us about the same environmental event?

According to Bregman, the auditory system solves this problem in two ways:

(1) by the use of *primitive processes of auditory grouping*, and
(2) by governing the listening process by *schemas* that incorporate *knowledge of familiar sounds*

4.2.1 Primitive auditory grouping

The primitive processes seem to represent *general acoustic properties*, i.e., properties that hold for a broad scala of auditory environments, and where the *meaning* that can be inferred from the acoustical signal is *not important* for the analysis of these *simple* signal properties. The auditory system *automatically* performs these *separate, local* analyses of signal components based on information about intensity, fluctuation pattern, direction of frequency transition, frequency modulation, spatial location information, etc. The incoming array of energy is broken down into a large number of separate analyses that are performed locally in time as well as frequency.

The next step is to *group* these signal components together so that each group is derived from the same environmental event. This grouping occurs along two dimensions, *in time* and *in frequency*, respectively known as *sequential integration* and *simultaneous integration*. These are *not* independent of one another. It is argued that these primitive analyses *and* grouping are performed *automatically*, without being influenced by attention, and without being influenced by learning. Therefore, these analyses are probably performed by *innate* mechanisms. This type of processing is often referred to as *input-* or *data-driven*, or *bottom-up* processing, because of the direction of the flow of information. The process of auditory grouping is guided by general perceptual organization principles, strongly related to Gestalt principles such as:

- *Proximity* in time or frequency (a form of *continuity*)
- *Similarity* on the basis of *global* properties such as pitch (F0), brightness, timbre, spatial location
- *Continuity* in pitch, formants and spatial location
- *Common fate*:
  - *Synchronicity of changes*: *Correlated* amplitude and frequency modulation (AM, FM) on a microscopic as well as a macroscopic level
  - Specifically, synchronous on- and offset
- *Exclusive allocation*: if one piece of sensory evidence has been assigned to one particular auditory stream, it can not, at the same time, contribute to another stream

Bregman (1990) argues that these properties can be interpreted from an *ecological perspective* since they can all be attributed to the *properties and physical limitations of the sources* that are *present in the auditory scene*. Before describing some of the experiments that show these primitive processes at work, the schema-based processes will be introduced.

4.2.2 Schema-based processing

The term *schema* is often used in cognitive science, and refers to a *cognitive structure* that can be either abstract or concrete. It can be seen as *some "control system" in the human brain that is sensitive to some frequently occurring pattern, either in the environments, in ourselves, or in how the two interact*. As these schemas refer to *specific* patterns, they are *domain specific*, and therefore not as general as the primitive processes. Every schema has its own methods for evaluating the sensory evidence to determine whether the pattern that it represents is present in the input. This pertains to the problem known as *assimilation*, which refers to the requirement that a schema has to be appropriately applied: it may only become (and stay) active *if and only if* the sensory evidence for which it stands is actually present. This is related to the requirement of *input-specificity* (Fodor, 1983).

Primitive grouping	Schema-based grouping
- General	- Domain-specific
- Simple, local analyses	- Global analyses
- Automatic, not influenced by attention	- Influenced by attention
- Innate	- Learned
- Data-driven (bottom-up)	- Hypothesis-driven (top-down)
- Preliminary linking of signal components	- Building possible descriptions based on grouping hypotheses

Table 4.1 Summary of the properties that distinguish the processes of primitive grouping from schema-based grouping, which are both involved in performing auditory scene analysis.

Other aspects that distinguish schemas from primitive processes is that they have a broader *temporal scope*, i.e., their analyses are more *global*. Furthermore, they are strongly influenced by stimulus-specific perceptual *learning*, and they are influenced by *attention*. In this context, the use of schemas during perception corresponds to the constructive approach, because there is a "search" for confirming stimulation in the auditory input matching the hypotheses that have been activated by bottom-up processes. It is therefore also known as *prediction-* or *hypothesis driven*, or *top-down* processing, because of the involvement of higher levels of information. The properties that characterize primitive and schema-based grouping are summarized in table 4.1.

In the case of acoustic stimuli the time dimension is of particular importance. Many of the patterns that schemas deal with extend over time and therefore represent a *temporal pattern*. Such schemas can therefore be interpreted as representing *dynamic systems*. When part of the evidence of a particular time-evolving pattern has already been detected and the schema representing this pattern has been activated, then this schema is in a state in which it is *primed* [40] to detect the following elements in this pattern.

4.2.3 ASA as a two-component process

By combining primitive and schema-based processes, Bregman (1990) describes the process of auditory scene analysis as a two-component process:

(1) First, the general, primitive processes perform their independent analyses on decomposed signal components, based on entirely *local* information. Auditory grouping then leads to *links* of different strengths between signal components and thus to the creation of *possible* organizations.
(2) The schema-based processes supplement these primitive processes and interpret the grouping hypotheses by using their knowledge of familiar patterns in the environment to reconstruct mental representations of individual auditory events, i.e., to form *descriptions* of signals stemming from single sources.

[40] Note that the use of the term *priming* refers to the *pre-activation* of the subsequent elements which is the result of the neural connections that exists between the elements that are represented in the schema, or dynamic system, as a consequence of previous experience, i.e., learning. This leads to a *bias*, or criterion-shift in the perception of these following elements, which does not necessarily implies the presence of a top-down feedback mechanism that directly alters the activation of new stimuli, or even feature-detectors that have not even been activated by the input. It merely stays activated when the new evidence is consistent with the activated schema, or hypothesis. If it is, it sooner accepted, i.e., less information is needed, than it would have been without this priming, and in case of a mismatch, the hypothesis is strongly inhibited. This issue will also be discussed in chapter 5, in the context of (psycholinguistic) models of spoken-word recognition and Signal Detection Theory (SDT, see appendixA).

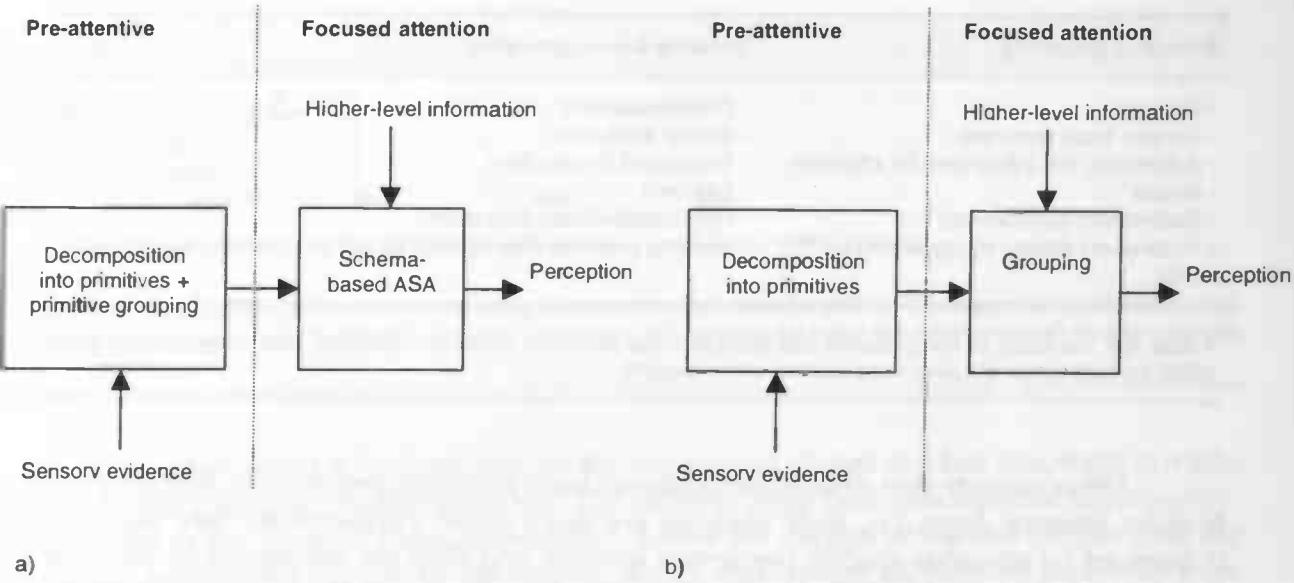


Figure 4.2 Simplified flow diagram for Bregman and Treisman's stages of perceptual processing. **a)** Bregman's two-component process where the pre-attentive stage decomposes the sensory evidence into primitives based on source properties which are preliminary linked such that schema-based knowledge can be applied to them in a focused attention stage. When needed, the preliminary linkages can be overcome. **b)** Treisman's two-stage processing sequence. Again, in the pre-attentive stage, the signal is decomposed into primitives which depends on the physical properties of the signal. In the focused attention stage these primitives are combined which is based on both stimulus input and knowledge derived from previous experience.

When comparing these two "stages" in perception with, for example, the stages of Treisman (1987), Bregman's stages differ in one important aspect, namely the influence of attention in the grouping process (see figure 4.2). Treisman suggests that perception takes place in two or more stages, where the first stage is the *pre-attentive* stage that decomposes the stimulus automatically and rapidly into a number of basic properties, called *primitives*. This analysis of objects into primitives corresponds again to bottom-up processing. Once primitives are extracted they are combined in the *focused attention* stage, which is *not automatic and requires conscious attention* (though it may be "influenced" or *describable* by Gestalt-like principles). According to Bregman, however, this grouping occurs automatically.

In addition, Treisman acknowledges that not only the nature of the units that make up "objects", but also our knowledge of the world influences perception. Therefore, top-down processing plays an important role. This effect comes *also* into play during the focused attention stage, in which primitives are being combined. According to Treisman, normal perception falls somewhere between the extremes solely from prior knowledge and construction solely from sensory data.

Before addressing the influence of attention in grouping, and the combined influence of bottom-up and top-down information, some experimental results that illustrate the different ASA processes at work will be described.

4.3 Auditory stream segregation

The general process of auditory scene analysis in which links are formed between parts of the sensory data, is known as *auditory stream segregation*, and has also been called *streaming*. A related effect, known as the *streaming effect*, refers to a phenomenon that can be created by presenting a short repeating loop of a sequence of alternating low-frequency (L) and high-frequency (H) tones: L-H-L-H-L-H-... Initially, subjects are able to follow this alternation pattern. After some time though, separate streams are formed; one containing the high-frequency components, -H—H—H..., the other containing the low-frequency components L—L—L-...

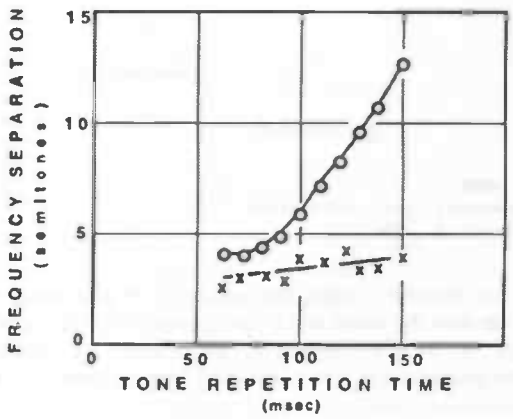


Figure 4.3 Temporal Coherence Boundary (o) and Fission Boundary (x) plotted as a function of the onset-to-onset time of adjacent tones. Higher times represent slower sequences. From van Noorden (1975).

This effect builds up over time as it depends on a *cumulation of evidence*. The reason why these repeating loops are used as stimuli is to *force* the grouping of *qualitatively similar* tones into separate streams. The streaming effect is not just a laboratory effect, but is argued to be influenced by the principles of stream segregation that are responsible for our normal perception of distinct auditory events, possibly consisting of multiple sounds, stemming from a single environmental source.

A number of factors, which are important for *sequential integration* in general, can have an influence on stream segregation. For example, when the time difference Δt between the (onset of the) different tones becomes smaller, it is more difficult to follow the alternation pattern, and segregation occurs sooner. Furthermore, when the distance Δf between the low- and high-frequency components is increased, the different streams also segregate faster.

4.3.1 The influence of attention

The streaming effect has been demonstrated with different kinds of stimuli. It seems though that the effect depends mainly on the task requirements (van Noorden, 1975). When subjects are asked to try to hear *all* tones as one stream, i.e., to hear them as a single *coherent* whole, there is a certain moment at which the auditory system is forced to segregate the tones into multiple streams. This is known as the *Temporal Coherence Boundary* (TCB), and is interpreted by Bregman (1990) as being the result of automatic, primitive organization processes.

However, when the task is to *select* a single stream, i.e., to focus *attention*, there is a point at which the auditory system is no longer capable of segregating the tones into different streams, and all tones are *fused* into one stream. This is known as the *Fission Boundary* (FB), and is interpreted to refer to the limits of our attentional capacity.

These different boundaries are shown in figure 4.3. It can be seen that between the TCB and the FB there is an increasing region of ambiguity as the frequency separation decreases and the onset-to-onset time differences increases. The subject can alternatively hear one coherent stream, or two segregated streams, *dependent on how attention is focused*. Bregman argues that primitive processes are not influenced by attention. Since these results show that the effect depends on the task instructions, they demonstrate that the effect *is* influenced by an attentional component.

Carlyon et al. (2000) also illustrated the effects of attention. They used a galloping rhythm, similar to the one used by van Noorden (1975), LHL-LHL-LHL-..., as depicted in figure 4.4. Again, when the frequency separation Δf is small and the sequence is played at a slow rate, listeners perceive the galloping rhythm corresponding to the repeating triplets of tones.

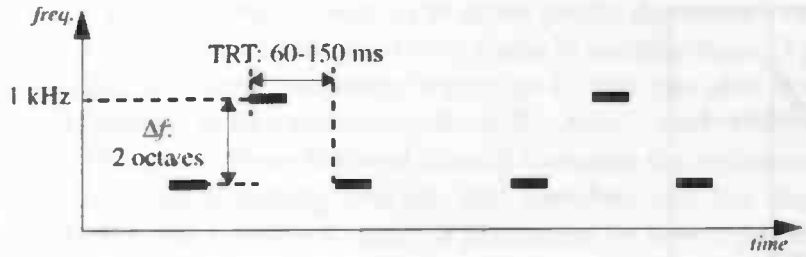


Figure 4.4 Stimulus configuration for the streaming experiments of van Noorden (1975). The sequences of alternating sinusoidal signals are presented with differing frequency separations (Δf) between the tones and differing overall repetition periods (TRT, or SOA).

However, when Δf is large and/or the sequence is played at a fast rate, there is a tendency for the galloping rhythm to be lost. The subjects again hear two separate streams, corresponding to the H and L tones respectively, where the rhythm of the one (in this case the L-stream) is twice the rhythm of the second. Again, this stream segregation builds up over time, so that for a given sequence, subjects hear a single stream (the galloping rhythm) at the beginning, and two streams at the end.

To investigate the influence of attention, Carlyon et al. (2000) presented similar sequences for 21 sec to the left ear, and subjects where given a competing task for the first 10 sec. After 10 sec, they were told to start making streaming judgments of the left ear sequence. The rationale behind this experiment is that if streaming is mediated by purely automatic, peripheral processes, then, after the first 10 sec, subjects should report the same amount of stream segregation as if they had been attending only to the sequence all along.

In contrast, if selective attention is involved in the build-up of streaming, then they should report a much lower amount of segregation, similar to what they would normally report at the beginning of the sequence. It was found that the competing-task condition during the first 10 sec indeed led to more "1 stream" judgments after 10 sec than were made without this competing task, dependent on whether the competing task required the subjects to attend to the left ear. Therefore, selective attention influenced the amount of stream segregation.

They also demonstrated that unilateral neglect patients, which has been shown to be an attentional deficit (e.g., Halligan and Marshall, 1993), also had a reduced build-up of stream segregation in the ear corresponding to the side of neglect. This reflects that damage to primarily cortical regions can affect streaming in a way that is not directly predictable by peripheral accounts of streaming, and that attention is an important determinant factor in the streaming process.

Nevertheless, these results are consistent with the influence of both rate and frequency separation. Probably, these aspects reflect the importance of the *continuity* of signal properties that are related to the physical limitations of a single, natural sound source. The primitiveness of these influences, in the sense of there being no influence of attention, remains an issue of discussion, and so far, at least questions the usefulness and relevance of distinguishing primitive and schema-based processes on this dimension.

4.3.2 Physiological breakdown versus functional accomplishment

Physiological explanations of the streaming effect typically refer to the limits of attention-shifting along a certain perceptual dimension. Alternatively, a break-down of pitch motion or jump detectors (*habituation*), could also be responsible for the inability to perceive the alternating sequence as a single coherent stream. According to Bregman, the different explanations do not exclude one another, since they describe the phenomenon from different per-

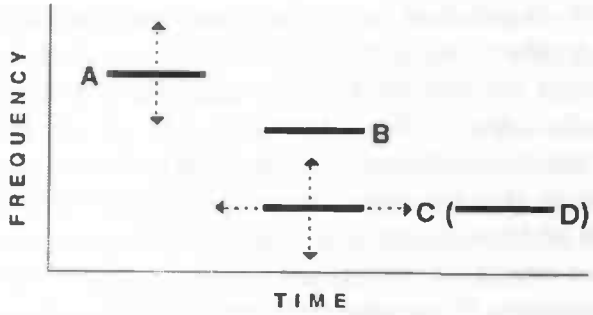


Figure 4.5 Stimuli used by Bregman and Pinker (1978). A, B, C, and D are pure tone components. Several ASA cues are manipulated resulting in competing and collaborating effects between cues.

spectives. A physiological breakdown can therefore be considered as a functional accomplishment, since both allow for the grouping of qualitatively similar auditory ele-ments.

There is another reason for considering the streaming effect as a functional accomplishment. Both the auditory and visual systems become inattentive, or in a sense turn off, when periodicity or symmetry is perceived over even a rather short time, and thus fail to extract critical information. This is especially true in the case of the auditory system.

Each natural sound source has its unique random pattern that differs in detail from all others to a degree detectable by the ear. Since this natural random pitch variation is a characteristic source property, it is important for perception. Cook (1999) even argues that if there were no imperfections in the source causing quasi-periodicity, perceptual fusion would not occur and a listener could neither identify its nature (*source identification*) nor hear parts (stream segregation). Nor could the listener recognize that there is more than one source per part. Thus, the auditory system is argued to be dependent upon these acoustic imperfections. Whether or not quasi-periodicity in the acoustic microstructure is really necessarily for perceptual fusion (*stream integration*) to occur, it is true that *all perceptual processing depends on changes in stimulation*, since such changes carry (most) information. It is therefore not excluded that the streaming effect results from the predictability of the stimuli used instead of resulting from a grouping process that is based on similarity in qualitative stimulus properties.

4.4 Competition and collaboration of cues

A task of the perceptual system is to reach an optimally coherent interpretation of the environmental scene. The grouping of different signal components as stemming from the same auditory event, can be influenced by different cues, which may or may not be consistent with one another. This can be illustrated by manipulating several cues for (sequential or simultaneous) integration. In figure 4.5, for example, a single tone A is alternated with a complex B-C. The ability of A to group with B (or to *capture* B) can be influenced by the frequency separation between the two: the smaller this Δf, the easier it is for A to capture B. This capturing can be strengthened by making the onset and/or offset of C different from that of B.

When B and C are in a *harmonic* relationship to one another, they are consistent as being partials of the same fundamental frequency F0. This strengthens the grouping of B with C. The *harmonicity principle* is an example of a cue that is very important for *simultaneous* integration. It can be explained by the fact that many natural sound producing sources are producing voiced sounds, where the harmonics of a particular F0 are simultaneously present. However, when there is an asynchrony between the onset and/or offset of these partials despite this harmonic relationship, B is released to group more easily with A. Whether

or not it actually groups with A is again partly dependent on the frequency separation between A and B. Furthermore, when there is another tone D present after the complex B-C, and within the same frequency region as C, this can also lead to the release of B to group with A, leading to two streams A-B—A-B—A-B-... and –C-D—C-D—C-D...

The use of these kinds of stimuli indicates the influence of competing and collaborating *forces* in the grouping process, and therefore also the importance of taking *context* into account. It is not possible to make a prediction on how a particular piece of sensory evidence will group at a certain instant in time, without taking all the current evidence, and the preceding *and* following context into account. Therefore, it is useful to interpret these grouping tendencies in terms of strengths or *forces of attraction* to the formation of possible groupings, without making definite *decisions* based on solely local information.

4.4.1 Exclusive allocation and capturing

Another important aspect that should be stressed, is that the existence of these competition effects is related to the Gestalt principle known as *exclusive allocation*: one piece of sensory evidence can *only* attribute to one perceptual grouping *at the same time*. It is this principle that makes the capturing of one tone out of a complex tone possible, even if this complex tone consists of a set of partials of the same F0. This capturing of a partial out of a complex of harmonics has also been shown for complex tones with more than two partials. However, the validity of this principle regarding sensory evidence is actually not demonstrated by these experiments. This is because the use of discrete tones partitions the stream into clearly identifiable *auditory elements* (or *auditory entities*) for which perceptual descriptions can be easily build. The experimental results may therefore only reflect competitive effects at the level of perceptual awareness, i.e., at the level where a consistent *overall* interpretation is made about the environmental scene.

4.4.2 Camouflage

The capturing effect has also been used to explain the phenomenon known as *camouflage*. When a certain familiar *tune* is played in the presence of other distractor tones it can be camouflaged. Dependent on the frequency relation between the distractors and the critical, relevant tones, the tune may or may not disappear in the pattern. Hearing out the tune therefore depends on the presence of distractor tones in the same frequency region of the relevant tones which can capture the distractors into the pattern. When the frequency separation is sufficiently large, this capturing will not occur. Bregman explains this effect as being the result of primitive grouping principles.

However, it is also very much influenced by the expectations of the listeners. When they are told which melody they should be able to hear, the frequency separation can be much less than it is without this knowledge. It therefore actually reflects schema-based segregation. It could be argued that the manipulation of cues important for primitive ASA can make schema-based segregation more difficult, since this is reflected by the importance of the frequency separation between the distractors and the critical tones. But, given the disparate results that are obtained when listeners know what to expect, it also depends on familiarity. It can therefore not be explained, or predicted, solely on the basis of acoustic cues. Schema-based segregation can overcome the primitive segregation tendencies.

4.4.3 Release of psychoacoustic dissonance

Another interesting effect is the *release of psychoacoustic dissonance* that occurs when two melodies are played synchronously. If two tones that form a dissonant pair (when played in isolation) are parts of different streams, because they belong to different melodies that are played (by the same instrument) at the same time but within clearly separated frequency regions, or within clearly separable streams, there is no sense of dissonance experienced by the listener. This is often used within polyphonic music styles. The "control of dissonance" can be achieved by ensuring that the simultaneous sounds are not assigned to the same auditory stream, thus by manipulating stream segregation cues. This allows more variety to be used in the musical composition. The interpretation of Bregman for the occurrence of this effect in terms of ASA is not an explicit part of music theory, but is consistent with how music theory deals with the control of dissonance.

4.5 Masking

A well-known *effect* in the field of psychoacoustics is that of the *masking* of a pure tone by noise or by other tones. Masking is defined as

(1) the *process* by which the threshold of audibility for one sound is raised by the presence of another (masking) sound, and
(2) the *amount* (in dB) by which the threshold of audibility of a sound is raised by the presence of another (masking) sound (American Standards Association, 1960).

The *masked threshold* is the sound pressure level (SPL) of a test sound, necessary to be just audible in the presence of a masker. This threshold, in all but a few special cases, always lies above threshold in quiet (absolute threshold); it is identical with threshold in quiet when the frequencies of the masker and the test sound are very different (Zwicker & Fastl, 1999).

If the masker is increased steadily, there is a continuous transition between an audible (unmasked) test tone and one that is totally masked. So, besides total masking, partial masking also occurs. Partial masking reduces the loudness, but does not mask the test tone completely. This effect often takes place in conversations.

Masking effects can be measured not only when masker and test sound are presented simultaneously, but also when they are not simultaneous. When the test sound is a short burst or sound impulse that is presented *before* the masker stimulus is switched on, this leads to a masking effect known as *pre-stimulus* masking (*premasking*), or *backward masking*. This effect is not very strong. Another form of non-simultaneous masking, where quite pronounced effects occur, results when the test sound is presented *after* the masker is switched off. This is known as *post-stimulus* masking (*postmasking*), or *forward masking*.

So, as in natural situations the presence of a sound stimulus rarely occurs in isolation, it is to be expected that all sorts of masking mentioned above also play an important role in dealing with the real auditory scene, and therefore with ASA. Off course, this also holds for *one* signal composed of multiple frequency components that are present at the same time, or follow each other in time. Though it would go to far to deal with all kinds of masker-stimulus combinations, their temporal effects, and their influence on loudness perception, some possible explanations for a limited number of masking phenomena will be discussed.

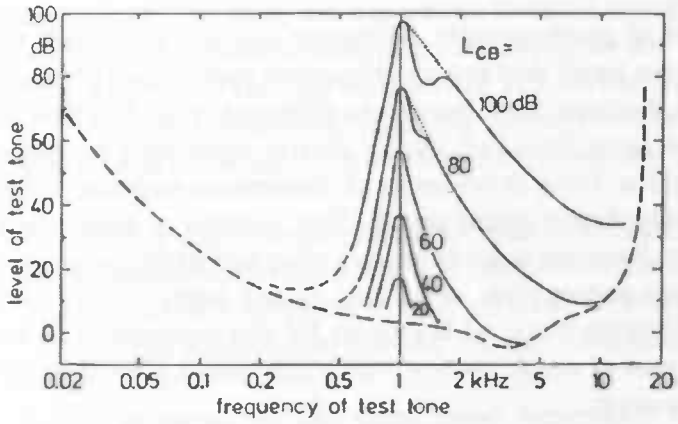


Figure 4.6 Masking of a pure tone by a narrow-band noise. Level of test tone just masked by critical-band wide noise with a centre frequency of 1 kHz and different levels as a function of the frequency of the test tone. The broken curve reflects threshold in quiet. The slopes on the high-frequency side are less steep, such that it is easier to mask a high-frequency component by the presence of a low-frequency component than vice versa. Note also the nonlinearity in the growth of masking with an increase in intensity at the high-frequency side. This results in an *upward spread of masking*: with an increase in intensity, the low frequency components become more and more effective in masking the higher ones. (From Zwicker and Fastl, 1999).

4.5.1 Masking of one tone by the presence of another tone

When the threshold of audibility of a tone with adjustable frequency is measured in the presence of a masker with fixed frequency and intensity, this masked threshold as a function of the frequency of the adjustable tone can be plotted in a *masking audiogram*. However, interpretations of such results are complicated by the presence of *beats* when the signal and masker are close together in frequency. These occur as a result of the changing phase relationship between the signal and masker, which causes the two to alternately reinforce and cancel each other. The rate of these resulting amplitude fluctuations is equal to the frequency difference between the two tones, which, if not too rapid, leads to audible loudness fluctuations (*beats*). As the task in a masking experiment is to judge whether or not a second tone is present, this qualitative change can be used as a cue to the presence of the signal even if the signal itself is not audible. Therefore masking experiments are often conducted with narrow bands of noise as either the signal or masker, because these have "built in" amplitude modulations and do not produce regular beats when added to a tone.

4.5.2 Masking of a tone by bands of noise

Results of an experiment where a pure tone is masked by a narrow-band noise with a certain center frequency are shown in figure 4.6. It can be seen that the slopes on the high-frequency side are less steep and depend to some extend on the level of the masker, i.e., the amount of masking grows nonlinearly on the high-frequency side. As the overall level is increased the low frequencies become more and more effective in masking the higher ones. The consequences of this *upward spread of masking* for speech, where each frequency component will have a masking effect on adjacent frequency components, is that it might result in a loss of audibility of the important information-carrying mid-to high-frequency components.

It also leads to an *asymmetry* in the masking of one tone by a second tone: it is easier to mask a tone by a second tone of lower frequency than by one of higher frequency. Furthermore, frequencies near the signal are more effective than frequencies farther removed from the signal.

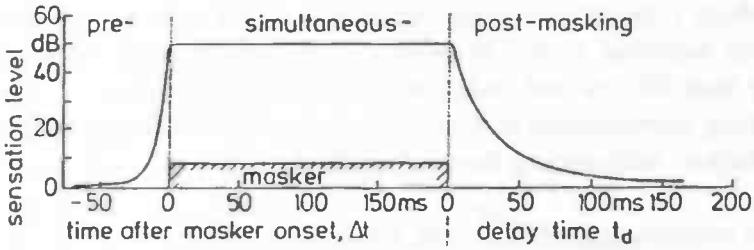


Figure 4.7 Schematic drawing to illustrate and characterize the regions within which *premasking, simulteaneous masking* and *postmasking* occur. Note that postmasking uses a different time origin than premasking and simultaneous masking. [From Zwicker and Fast, 1999].

4.5.3 The critical band

These and other masking phenomena have led researchers to use the concept of a *critical band* (introduced by Fletcher, 1940). It has been suggested that different frequencies produce their maximal effects at different locations along the basilar membrane, so that each location responds only to a limited range of frequencies. The effective range to which a given location (or filter) responds is its critical band. This critical band should not be considered as a rigid rectangular filter, but as a continuous series of overlapping bands with a certain range and with sloping edges.

It has been shown that only a narrow band of frequencies surrounding the tone contribute to the masking of the tone. Broadening the noise band outside this critical bandwidth does not elevate masking threshold. There have been a large number of different experiments that have shown that listener's responses to complex sounds differ according to whether the stimuli fall within or outside the critical band. Furthermore, these different experiments give remarkably similar estimates both of the absolute width of the critical band and of the way it varies as a function of frequency. Note however, that the critical band should be interpreted as a concept that is useful for summarizing many experimental effects, and therefore refers to a convenient *description* at a higher level of abstraction.

Masking has been *explained* in terms of BM displacement and, in particular, the neural firing pattern associated with this, where both the neural excitation pattern for a certain group of neurons (the *amount* of neural recruitment), and the fine-time characteristics of neural firings at several places in the auditory system as a result of phase-locking (the *temporal pattern*) can play a role (Green, 1976; Moore, 1977; Allen, 1995; Zwicker & Fastl, 1999). It has also been interpreted in terms of IHC excitation patterns (Neely & Dai, 2000), which are off course also linked with the excitation of the auditory nerve fibers to which they are connected.

4.5.4 Temporal effects

The transmission of information in, for example, speech implies a strong temporal structure of the sound. Loud sounds are followed by faint sounds and vice versa. In chapter 2 it has already been mentioned that vowels generally represent the loudest parts whereas consonants are relatively faint (i.e., vowels are more sonorant). Therefore, a plosive consonant is a typical example of a sound that is often masked by a preceding loud vowel. This is a result of the temporal effects of masking that characterizes our hearing system.

To measure these effects quantitatively, maskers of limited duration are presented and masking effects are tested with short test-tone bursts or short pulses. Further, the short signal is shifted in time relative to the masker (see figure4.7). It can be seen that three different temporal regions of masking relative to the presentation of the masker stimulus can be

differentiated. The time during which forward masking can be measured is relatively short (about 20 ms), whereas backward masking is the dominant non-simultaneous temporal masking effect as it can last longer than 100 ms and ends after about a 200 ms delay.

The effect of forward masking corresponds to a decay in the effect of the masker and is more or less expected. Possible factors influencing forward masking are:

(1) a reduction in sensitivity of recently stimulated cells, i.e., *auditory fatigue*,[41] or
(2) a persistence in the pattern of neural activity evoked by the masker, or
(3) a result of a temporal overlap of cochlear responses because the vibrations evoked by the first stimulus on the BM will not have completely died away (Duifhuis, 1973).

Backward masking, on the other hand, represents the fact that *each sensation does not exist instantaneously, but requires a build-up over time to be perceived* (see also section 4.6.2.2, and section 5.4.3). It may also be related to temporal overlap of BM recruitment.

These explanations all refer to peripheral processes. However, to account for longer Δts it seems that higher neural levels, and more central processes are involved. This is strengthened by the finding that backward and forward masking effects are also observed in dichotic listening: frequency components presented at one ear can have a masking effect on frequency components at the other ear.[42]
Both absolute threshold and masked threshold depend on:

(1) the *duration* of the test sound, and/or
(2) on the *repetition rate* in the case of repeated short test sounds, which are usually tone bursts.[43]

Both dependencies lead to identical results for absolute threshold and for masked threshold:

- For durations *longer* than 200 ms (or for repetition rates corresponding to 5 Hz) the test-tone threshold is constant, corresponding to that of long-lasting sounds.
- For durations *shorter* than 200 ms, absolute threshold and masked threshold increase with decreasing duration at a rate of about 10 dB per decade.

---

[41] A distinction is made between *auditory fatigue* and *auditory adaptation*. Auditory adaptation has as its essential feature the process of *equilibration*. The response of a receptor to a steady stimulus declines as a function of time until it reaches a steady level at which the energy expended by the receptor is just balanced by the metabolic energy which becomes available to sustain it. Adaptation typically occurs *during presentation* (for example during simultaneous masking) and parallels with neural adaptation. Auditory fatigue is concerned with the effect of an excessive stimulus upon a small and finite group of receptors, namely those which are normally brought into activity at near threshold intensities. It is therefore thought to be connected with hair cell response changes, and occurs *post-stimulatory*. It results in a *temporal threshold-shift* (TTS). However, the recovery process may be quite rapid. As fatigue and adaptation often do not refer to completely independent physiological processes, these processes and/or their effects on perception, are not always easy to distinguish.

[42] Though this might appear to be strange, it should be remembered that masking is determined within an experimental context. The masking effect therefore depends very much on the task instructions, and actually reflects the subject's *interpretation* of the *total* auditory scene. Presenting stimuli to both ears, increases the amount of evidence on which a subject's response is based. This evidence might include information that favors the integration of information of both ears as representing one event. It is difficult to distinguish between these different bases of the masking effect by just looking at the subject's response.

[43] Since a reduction of the duration of such tone bursts produces a widening of its spectrum (compare with impulses), this limits the shortest usable duration to a value at which the spectral width corresponds to that of the critical band. A quick rise time that occurs with a narrow spectrum is produced by Gaussian rise and fall. Therefore Gaussian-shaped tone bursts are usually used to measure temporal effects.

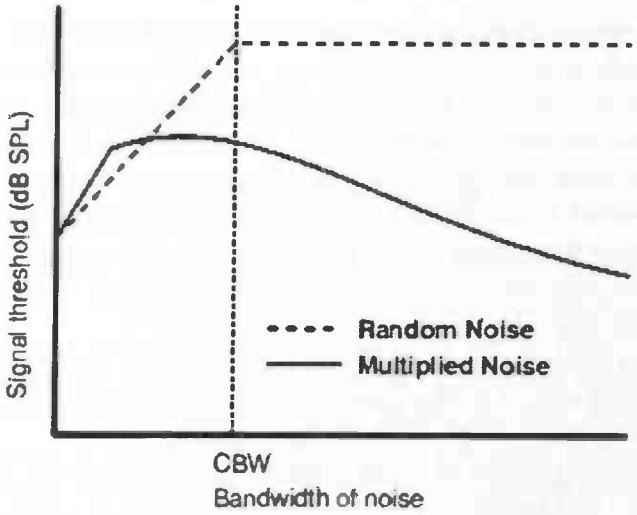


Figure 4.8. Detection thresholds for a 1 kHz signal centered in either random noise or noise amplitude modulated at a slow, irregular rate, as a function of the noise bandwidth. Masker spectrum level is held constant. [From Hall et al., 1984].

It is therefore often assumed that the auditory system integrates the sound intensity over a period of 200 ms. There are several models regarding such a *temporal integration window*, but that this should be a fixed and long time window is doubted (Dai et al., 2000; Plack & White, 2000). Some of these models will be discussed within section 4.6.5. It can be mentioned here that, consistent with a *decrease* in temporal *resolution* at higher levels of processing, it has often been found that there is an *increase* in the length of a possible temporal *integration* window at higher levels of processing (see chapter 3). The length of 200 ms probably corresponds to processing at the level of the auditory cortex, and it is therefore not surprising that in psychoacoustical experiments where the responses are the outcome of the whole system often reflect this value. In chapter 5, section 5.3.4, some more findings regarding this 200 ms interval will be described.

4.5.5 Release of masking

There are some interesting phenomena where the effect of masking is released by the presence of cues that also seem to be important for the fusion into auditory streams.

4.5.5.1 Comodulation Masking Release

One such phenomenon is known as *Comodulation Masking Release* (CMR), discovered by Hall et al. (1984). In their experiment, detection thresholds for a 400 ms, 1 kHz pure tone were measured when masked by different types of noise. When the masker was a band of random noise centered on the frequency of the target tone, then increasing the bandwidth of the noise, while holding the spectral level constant, led to an increase in detection threshold. After a certain width was reached, the interference got no worse, which can be explained by the concept of a critical band, i.e., only the energy in a spectral band adjacent in frequency to a given target is effective in masking it.

However, when the masker was not a simple noise band, but had been amplitude modulated in a random pattern, and when the same type of AM noise was added while increasing the bandwidth, this led again to an increase of the amount of masking up to the limit of the critical band. But as the bandwidth of the masker was made wider, the masking

began to *fall* in effectiveness. So, adding more energy to the masker reduced its ability to mask. These results are depicted in figure 4.8.

This effect can be explained as follows. As has been pointed out by Haggard, Hall and Fernandes (1984), correlated AM across frequencies serves as an important cue for ASA. This is because *many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories are highly correlated across frequency*. This is true of speech, of interfering noise such as "cafeteria" noise, and many other kinds of environmental stimuli. For such stimuli, the auditory system uses *across-frequency analysis of temporal modulation patterns* to help register and differentiate between acoustical sources.

The effects of AM are, however, also explainable in terms of the so-called "peek" theory: the reason that it is easier to hear a target sound in the presence of another sound when both are AM modulated in a different pattern, may simply reflect the fact the auditory system is getting a better peek at the target component at those instants of time at which the others have been reduced in intensity by the modulation. This could be the result of a better *local signal-to-noise ratio* (SNR) (see also Allen, 1996).

It has also been argued that those parts of the stimulus which occur during a minimum of the modulation envelope are given larger weights. Such a strategy is called *selective listening*, or *cued listening* (see also Nelken et al., 2000). But, how can these theories explain that the "peeks" get more effective when the bandwidth of the masker increases? One possibility might be that adding more evidence *outside* the critical band, leads to a further improvement of the local SNR by such a weight- adjustment process as it is based on more *information*, whereas adding more energy within the critical band does not add additional information. According to this interpretation, the concept of a critical bandwidth seems to be related to that of informational content. Indeed, many researchers see the critical band as representing roughly independent channels as far as their information carrying potential is concerned (e.g., Allen, 1996, 2000). Note that this has nothing to do with seeing them as *physically* independent channels, or filters!

Interestingly, the CMR does not seem to work with frequency modulations (FM). This is an important finding, because one of the ASA grouping principles is that of common fate which refers to correlated changes in AM *as well as FM*. For example, it has been shown that parallel frequency modulation of subsets of partials leads to stronger integrative effects. However, this could be attributable solely to the harmonicity principle, because with these stimuli the harmonic relationship is constantly retained. The CMR data suggest that this is indeed a correct interpretation. This is strengthened by the finding that when different frequency components move in parallel according to equal frequency *differences* (instead of the same frequency *ratios* as is the case with harmonics of the same F0), *multiple* sources are heard. Not surprisingly, there is no integration of these linear parallel frequency changes, but they tend to segregate, likely because natural sources do not produce such signals. Therefore, it appears that there is no general *common fate in FM* principle, which explains the absence of CMR with FM modulated noise-bands.

4.5.5.2 Masking release in forward masking

Another masking effect in which widening the masker bandwidth reduces the amount of masking, is found in forward masking where a loud noise burst masker precedes a fainter pure-tone target, and the masker is centered at the same frequency as the target. An increase in bandwidth (outside the critical band), again leads to a decrease in masking. This could be due to the fact that the additional energy makes the masker more distinctive, because a wide-band noise does not sound as much like the target as a narrow band of noise does. So, in ASA terms this would be interpreted as follows: *The fusion of the various spectral regions in the wide-band noise create a larger perceptual entity that has global properties that are different from the properties of its parts.*

Furthermore, a wider noise band again provides *more information* regarding the simultaneous offset of energy across the spectrum and therefore provide stronger evidence for the masker going off.

4.5.5.3 *Spatial release from masking*

Masking can also be released when target and masker are perceived to be at different locations, or just through the addition of localization-related evidence, known as *spatial release from masking*. This can be illustrated by the *binaural masking level difference* (BMLD). When, for example, a tone and a noise are presented binaurally, and the intensity of the noise is increased until the tone is no longer audible, changing the phase of the tone in one ear will make the tone audible again. The intensity of the noise has to be further increased in order to mask the tone again. The same result can be achieved by first presenting both tone and noise to one ear, and increasing the intensity of the noise so that it just masks the tone. Then, presenting only the noise to *both* ears, results in the tone becoming audible again. So again, providing more information leads to a reduction in masking.

It is important to note hear that in the case of a phase difference, this difference can be as much as 180°. This difference is greater than would occur in natural listening situations, and therefore the localization of the phase-reversed sound would be diffuse. This seems to imply that it is not required for the target and mask to be clearly localized, all that matters is *discrimination*. The between-ear phase comparison just has to lead to *different* results for target and masker.

4.5.5.4 *Masking and fusion*

It seems that the same factors that segregate one acoustic component from others also prevent that component from being masked. In both masking and fusion the acoustic component that is being masked, or being absorbed into a stream that contains other simultaneous components, gives up its perceptual identity in favor of contributing to the global features of the whole sound. The acoustic relations affecting both phenomena are, in a natural listening environment, useful for deciding whether different spectral components have arisen from the same acoustic event. *In the absence of such acoustic cues for source segregation or auditory grouping, the auditory system is biased to integrate the perceptual evidence.*

Furthermore, the findings on the release of masking illustrates that masking is definitely not exclusively influenced by peripheral processes, or physiological limitations. It therefore does not necessarily reflect the frequency resolution power of the BM as it has often been interpreted. Off course, it is important to understand the limitations at the peripheral level, for these also determine the limits of the rest of the system. But, results from masking experiments are probably not the most suitable. The task requirements do not allow an unambiguous interpretation of the results, and therefore do not exclusively measure the limitations they were originally supposed to measure. It is clear that more central processes also have an important influence, and masking is therefore *the perceptual result* that is determined by processing at *all* levels within the *whole* system. In general, it can be stated that the more *information* is available, the more likely it is to perform auditory grouping, and the more likely it is to find a release in masking. Different sources of information are processed at different levels of processing, and these together determine the perceptual outcome.

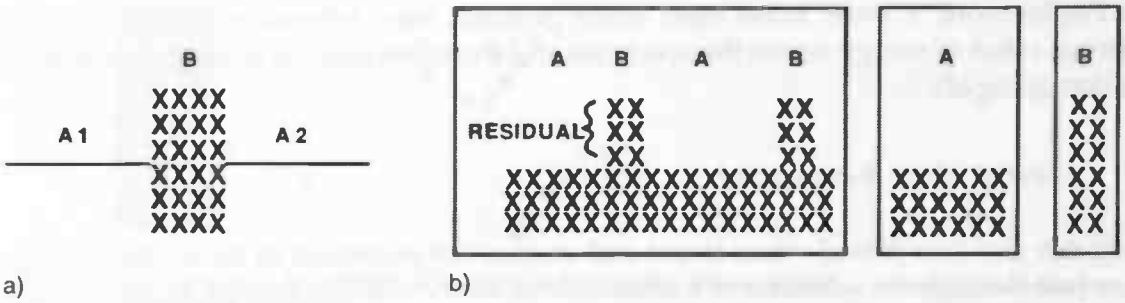


Figure 4.9 The "continuity illusion". (a) A1 and A2 are the parts of a longer tone A that is interrupted by a louder sound B. (b) An alternating pattern of low-frequency small-band noise A with a broad-band noise B leading to the formation of a high-frequency residual of B and a perceived continuous of the low-band noise A.

4.6 The "continuity illusion"

A well-know illusion that occurs in several perceptual modalities, is the illusion of perceptual continuity, or perceptual filling-in. In audition, there are also a number of perceptual phenomena that reflect this *continuity illusion*. Such phenomena seem to be unexplainable by bottom-up processes alone. As the perceived content of the sound is in some sense incorrect or different from what was actually present, this implicates that the result of perceptual organization has been additionally influenced by high-level *biases* based on the wider context of the stimulus or other information.

The effect can be illustrated by presenting a certain sine tone A, interrupted by a band of noise B (see figure 4.9a). When the noise energy is low relative to the tone, the sequence is heard as an alternation between the two. However, when the energy of the tone is decreased, or the noise energy increased, the perception changes to a steady, continuous tone to which short tone bursts have been *added*, rather than hearing the noise burst as *replacing* the tone A. The excitation due to the noise in the portions of the inner ear responding to the sine tone has become sufficiently high that it is impossible for the auditory system to "decide" whether or not the excitation is the result of the sine tone or the noise burst. So, as a result of the increasing energy of the noise burst relative to the tone, it has become increasingly ambiguous.

This effect also occurs when two tone burst are alternated, for example a low-frequency narrow band of noise A and a broad-band noise B (see figure 4.9b). Again, the narrow band of noise A is heard as continuous, whereas the broad-band noise B is now heard as a *high*-frequency band of noise that is alternately added to the continuous noise band A. So, the presence of A before and after B leaves only a high-frequency *residue* of B. This leaving of a residue behind is an important characteristic of primitive ASA, because it illustrates that the sensory evidence has been *partitioned*. Such partitioning of evidence into auditory streams has a *symmetric facilitating* effect on hearing *both* streams. Schema-based ASA processes, on the other hand, are characterized by taking from the sensory evidence what they need, *without* leaving a residue behind. So, schemas *select* from the sensory information the evidence corresponding to the temporal patterns they represent, and leave it there for other schemas to be used. As a result, when a certain signal is *recognized* in the presence of background noise, this does *not* lead to a facilitation of recognizing what is present in the background (Bregman, 1990). This distinguishes primitive from schema-based ASA, and will be further discussed in later sections.

Returning to the continuity illusion, it should be noted that the listener truly "hears" the tone, i.e., by the time the perception reaches levels of conscious introspection, there is no distinction between a percept based on "direct" acoustic evidence and one merely inferred from the context.

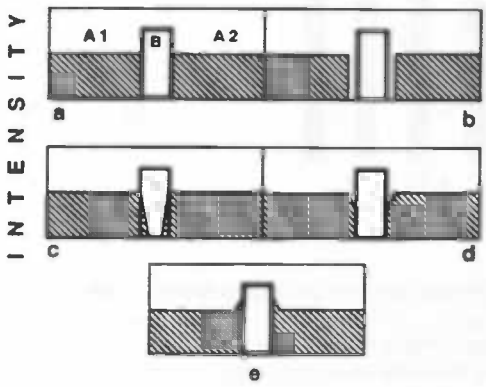


Figure 4.10 The effect of the boundary between sounds on the continuity illusion. (a) B is much louder than A and in the same frequency region. It masks the offset and subsequent onset of A. (b) There is a silence between A and B so that the offset and onset of B are audible. (c) B is not loud enough to mask the offset and subsequent onset of A. (d) A is perceptually dropping in intensity just at the onset of B and rising just after B's offset. (e) A is perceptibly rising in intensity just before B and dropping just after B.

It can be argued though, that this effect actually does not illustrate an illusion. In the case of the *sine tone - noise burst* alternation, an increase in the SNR, leads to an increase in ambiguity, especially because of the randomness of the noise, i.e., the *unpredictability* of its phase. This means that there is no conclusive way to establish if the tone was on or off during the noise. It can therefore be interpreted as a *justified, systematic bias* towards a particular kind of interpretation in a genuinely ambiguous situation.

The same effect has also been found with parts of speech utterances, which indicates that it occurs with both primitive and schema-based ASA. When parts of a sentence are interrupted with periods of silence which are then filled up with relatively loud noise bursts, the sentence is heard as continuous. (Parts of) the phonemes, that in fact are *not present*, are perceptually *"restored"* by the listener. This is not the case when the sentence is presented with silent interval interruptions. This effect has been known as the *picket-fence effect*, or as the *phoneme restoration (PR) effect*. The PR effect will be dealt with in more detail in chapter 5 (sections 5.4.3.3 and 5.6.5).

In what follows the continuity illusion will be considered in more detail, since it nicely illustrates many aspects that are related to ASA. Bregman has extensively tested under which conditions the continuity illusion occurs, and which principles, or possible mechanisms, can account for this "illusion". In general, it can be stated that, for the continuity illusion to occur, all the sensory evidence should *not exclude* the fact that A actually continued during the interruption and has not either gone off altogether or turned into B. Therefore, the following factors determine whether or not the illusion occurs.

4.6.1 Masking of discontinuities

There should be no evidence that B is actually covering up a silence between A1 and A2 rather than the continuation in A. So, there should be no evidence that A actually shuts off when B starts, or turns on again when B finishes (as in figure 4.10b and d). This is a useful strategy in dealing with the environment, because normally *the probability that a sound event, which is produced by another source, begins at exactly the same time that another sound event ends, is very small*. So, when A is already present and there is another source of energy introduced by the onset of B, it is normally "safe" for the auditory system to assume that A continues, unless there is evidence *before* the onset of B that A has finished. This has been called the *old-plus-new* heuristic, and also explains the residue effect of the two alternated narrow- and broad-band noises. Therefore, *the beginning and ending of an interrupting sound will define its own temporal boundaries but not the boundaries of the sound(s) it interrupts.*

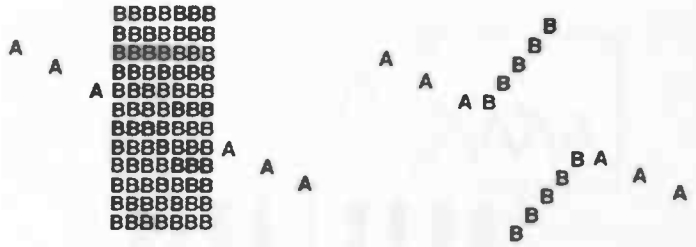


Figure 4.11 Illustration of the principle of simplicity in the selection of energy from an interrupting sound to complete the signal that occurs before and after it.

Furthermore, the continuity illusion seems to refer in particular to the *masking of discontinuities* such as masked on- and/or offsets.

4.6.2 Sufficiency of evidence as reflected in neural activity

During B, some of the neural activity in the auditory system should be indistinguishable from activity that would have occurred if A had actually continued. So, if there is contextual evidence that a sound may be present at a given time, and if the perceptual units of the auditory system stimulated by a louder sound include those which would be stimulated by the anticipated fainter sound, then the fainter sound may be heard as present (Warren, 1982). This is one of the reasons why the continuity illusion is sometimes explained, in functional terms, as a "compensation for masking" effect.

When for example a 1 kHz tone A is alternated with a broad-band noise B that is band-stop filtered for frequencies around 1 kHz, this leads to a reduced masking effect on A. Unless the loudness of tone A is very low, the noise B does not allow tone A to be heard as continuous. Therefore, it is important that the spectral content, that is the peripheral neural activity of B, must be as such *to sustain the hypothesis* that A is indeed present in the larger mixture. Note that the requirement applies to neural activity *not* to physical acoustic energy! A strong stimulus at a frequency far away from the frequency of the tone may spread its (mechanical or neural) activation to neural pathways far removed from those that are primarily tuned, but nevertheless might also activate the pathways that normally respond to A.

4.6.2.1 *The simplicity principle*

The parts that are selected from the interrupting sound B should *not only* be *consistent* with the hypothesis that has been activated by the parts of the sound A that occur on both sides, but it also seems to be heard as the *simplest* sound that is consistent with those parts. When for example, the A-parts on both side of B follow a trajectory, this trajectory will be filled in (see figure 4.11, left part), presumably because it is the simplest continuation of the two parts. Other consistent hypotheses, like hearing two parallel glides B in-between as illustrated in figure 4.11 (right part), are not heard. Normally, this pattern leads to two so-called bouncing percepts: a high-pitched glide first descending and then ascending, accompanied by a low-pitched glide that first ascends and then descends. Not only is this pattern not heard, it is also considered to be very unlikely that the auditory system creates it as a possible hypothesis among many other possible hypotheses during the presence of the noise, since there is really no basis for doing this.

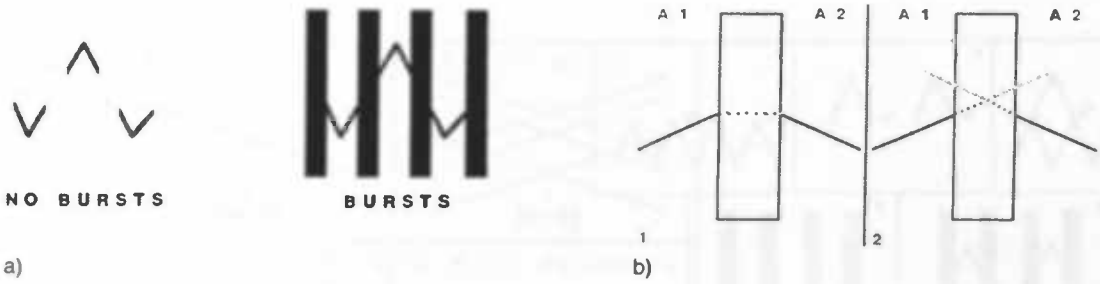


Figure 4.12 (a) Tonal glides of the type used by Dannenbring (1976). Left: the stimulus with gaps. Subjects clearly hear the interruptions with silence. Right: the stimulus when the gaps are filled with noise. In this case, subjects hear the glides as continuing behind the noise. (b) Illustration of two possible mechanisms for the restoration of glides through noise: (1) interpolation, and (2) extrapolation. The experimental evidence supports the former.

4.6.2.2 *Retroactive effects: interpolation versus extrapolation*

The previous example should therefore not be interpreted as evidence for a *trajectory following* process. For instance, with the pattern of stimuli that are depicted in figure 4.12a, the glides are again perceived to continue behind the interrupting noises as a continuous glide that alternately descends and ascends. Though the trajectory in-between appears to be filled in, this does *not* seem to be the result from a trajectory-following prediction process.

When the "points of return" of such descending and ascending glides are interrupted by noise (see figure 4.12b), the perceived pitch[44] at these return points is not the same as would have been when the trajectories were extrapolated (Dannenbring, 1976). Although the signal is perceived as continuous, the trajectory under the noise bursts is again as simple as possible: the highest frequency points that remained (after the peaks were deleted at both sides) are connected, leading to an alternately descending and ascending glide with less pronounced peaks, i.e., with relatively lower pitches at the peaks after the ascending part, and relatively higher pitches after the descending part.[45]

This is important, because it illustrates that the evidence that is present *after* the noise burst, determines how the sound *before* it is continued, which shows that there are *retroactive* effects in perception. This is not surprising, since our perceptual awareness is delayed relative to the presence of the stimulating physical evidence (see also section 5.4.3).

In addition, it indicates that there does *not* exist a *prediction mechanism* based on the previous direction of change in frequency that extrapolates frequency components along the same trajectory. This seems to be a valid strategy, because *natural sounds always change continuously*. For instance, at a microscopic level, the pitch contour is in a constant flux (see also section 4.3.2). It is therefore not very useful to assume a continuation of a certain direction of change, but rather to be prepared to perceive a change in direction.

In itself, this does not imply that the direction of frequency change is totally unimportant. When at a certain instant in time, one partial moves in a certain direction, the other partials of the same F0 change in the same direction (parallel on a log-scale). As has been mentioned earlier though, the harmonicity relation between the partials remains also. In addition, it is very difficult to estimate the amount and direction of the *instantaneous* frequency change in noisy conditions. Therefore, in general, frequency proximity and harmonicity and not common fate in FM seem to be the determinant factors for ASA.

[44] The perceived pitch can be determined by a pitch adjustment procedure, where subjects adjust a second ascending frequency glide pattern such that its highest frequency matches the highest pitch they hear behind the noise at the return point.

[45] Besides the simplicity principle, this suggests that there could be a *bias* towards perceiving smoother turnaround points, as this corresponds to a more "natural" stimulus.

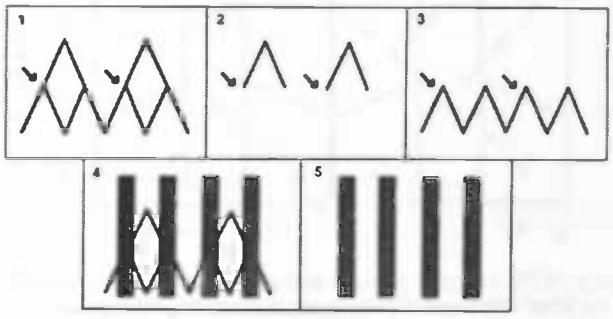


Figure 4.13 Pattern of glides used by Steiger (1980) to study the effects of perceptual restoration on stream organization. The pattern of glides in (1) branches to form the two streams represented in (2) and (3). When the "decision nodes" are obscured by a 20 ms noise burst as shown in (4), the same streams are formed, accompanied by a third stream representing the noise burst (5). The interrupting noise does not influence the grouping at both sides of the noise.

4.6.3 Evidence for source continuity as reflected in A1-A2 grouping

Another factor determining the occurrence of the continuity illusion is that there should be evidence that A1 and A2 actually came from the *same source*. The rules for sequential grouping would normally put them into the *same stream*, even if they had been separated by a silence instead of by B. In other words, the continuity illusion depends on the inter-pretation that one sound has *interrupted* another which means that A1 and A2 have to be interpreted as parts of the same sound, but not as part of the same sound as B. So, continuity through interruption is a special form of stream *integration* that occurs when it is plausible to interpret A1 and A2 not only as part of the same stream, but as a single continuous event within that stream. Therefore the requirements for the continuity illusion *include* those for stream integration.

To illustrate, when a pattern of glides branches apart at so-called "decision nodes" (see part 1 of figure 4.13), the pattern segregates into the two streams shown in part 2 and 3 of this figure (Steiger, 1980). The absence or presence of a 20 ms noise burst substituting the decision nodes leads to the same percept. In the presence of noise, these streams are accompanied by a third stream containing the noise burst (shown in part 5).

Again, later stimulus information after the decision nodes are incorporated in the same streams on the basis of frequency proximity and not on the basis of good continuation of glide trajectories. Therefore, the restoration is not performed by extrapolating the glide from the segment prior to the gap, but is based on information on both sides of the gap, obtained within a more global temporal scope.

In another experiment, performed by Tougas & Bregman (1985b), the stimuli shown in figure 4.14 were used. They also demonstrated that the perception of continuity through noise does not affect the organization into streams. Within the gliding patterns, the content of interval *I* was varied: it was either filled by glides, noise or silence. The stream organizations were similar regardless of the acoustic content of the interval:

- Pattern A led to a "bouncing" percept: a higher descending glide that at the midpoint ascends again, accompanied by a lower ascending glide that descends again. This is based on frequency proximity, i.e., the grouping of glides that stay in the same frequency region is favored.
- With pattern C, a "crossing" percept is perceived: a descending pure glide crossing an ascending rich glide, as the ascending glides are fused by the auditory system based on their harmonic relationship.
- Pattern B leads to a bouncing percept, because the harmonicity principle does not favor any interpretation, and therefore only frequency proximity seems to be important.

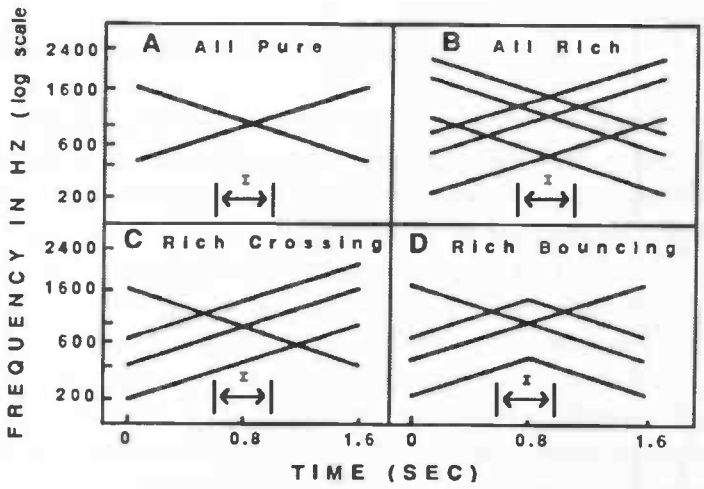


Figure 4.14 The four glide patterns (A-D) used by Tougas and Bregman (1985b). Interval I was either filled by glides (as shown), noise or silence. Most often, these stimuli led to the following percepts, where the determinant grouping principle is mentioned between brackets: (A) Bouncing percept (frequency proximity), (B) Bouncing percept (frequency proximity, no favoring by harmonicity), (C) Crossing percept (harmonicity), (D) Bouncing (frequency proximity and harmonicity).

- Finally, pattern D leads to a strong bouncing percept as both the harmonic relations between the first ascending and then descending rich glide, and frequency proximity favor this interpretation.

The interpretation of the bouncing percept in favor of the crossing percept, and the absence of a predictive trajectory-following percept has been much more extensively studied by Bregman and others. For instance, it has been shown under the following circumstances (see figure 4.15):

- discrete steady tones,
- interrupted glides aligned in the direction of the trajectory (even with an overlap at the cross-over point such that the constant-length units had to be broken to perceive the bouncing percept),
- adding rhythm as an extra cue either favoring the bouncing or crossing percept, etc.

Frequency proximity is always favored over trajectory-following. The only manipulation that favors the trajectory is by timbre, or harmonic relationship, but off course this should not be interpreted as evidence for the following of a trajectory and neither as evidence against the influence of frequency proximity. Furthermore, although frequency proximity is a form of the continuity principle, continuity in itself is not enough to predict what will happen, because the trajectory is *also* a form of continuation.[46]

According to Bregman (1990), the results regarding masking and the continuity illusion might be functionally explained as follows: First, auditory streams are formed as a result of heuristics for grouping. Then, in response to cues for masking (resulting from the presence of noise), a "mechanism is activated" to interpolate in the gaps between the events belonging to the already formed streams which can not affect the formation of such streams.

[46] The strength of the frequency proximity principle might be interpreted as reflecting a pitch-following process, as it is the pitch contour that can only change within certain time limits in a relatively small frequency region, if it is to reflect the pitch produced by a natural source. So, the more general term continuity, in my opinion, should only be used when referring to the continuity that results from the physical limitations of the sound source (more specifically, articulatory continuity, see also section 2.4), and not necessarily to spectrally visible signal components

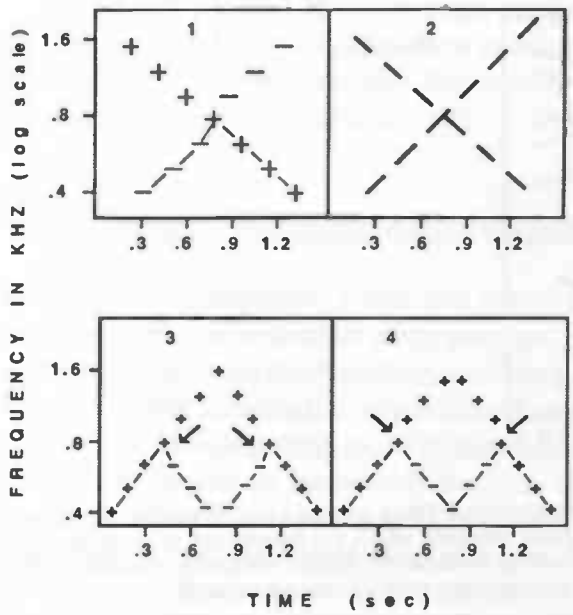


Figure 4.15 Competition between segregation by frequency region (the broken connecting lines) and by the following of a trajectory (pluses). (1) the crossing of an ascending and a descending pattern of tones, (2) crossing trajectories in which the individual tones are short glides, that are aligned in the trajectory direction and overlap in time, (3) the rhythm favors the trajectory, (4) the rhythm favors segregation by frequency range. (Tougas & Bregman, 1985a). Despite these different manipulations, the patterns are all interpreted as a bouncing percept in favor of a crossing percept. Thus, the perceptual interpretation is a higher descending glide that at the midpoint ascends again, accompanied by a lower ascending glide that descends again.

It seems plausible that in natural listening situations, when several sounds are usually occurring at the same time, the momentarily presence of a relatively loud sound should not lead to hearing different sounds than before the interruption when this loud sound has disappeared. In addition, as it is not desirable to perceive "chimeric" sounds created by grouping the earlier part of one signal with the later part of another, the stream organization processes must not be changed by the processes that restore continuity. However, the presence of A2, appearing as soon as B has disappeared, is *essential* for per-ceiving continuity during B.

As has been mentioned, the requirements of the continuity illusion *include* those for stream segregation. This is related to the fact that there are two aspects to the partitioning of a mixture of environmental sounds:

(1) One is to *factor out of the mixture the acoustic energy that arose from the same source*, for example the grouping of the sounds of a violin into a single stream, distinct from all co-occurring instrumental sounds. Therefore, interruption by a louder sound, leading to continuity, should not affect the perceived groupings.
(2) Another is to *partition the energy of that source into separate events*, for example hearing each note within the "violin stream" as a distinct event. Therefore, the rules of grouping *should* affect the restoration of the perceptual continuity.

Only when both levels of grouping are accomplished, the rhythmic pattern of the notes played by the violin can be appreciated. If the notes are not part of the same stream, but strongly segregated into different streams, the "intended" rhythm will not be heard. At the same time, unless there are distinct events starting at different times, there will be no rhythm in the stream.

4.6.4 No gradual transformation

A final factor influencing the continuity illusion is that the transformation from A to B and back again should not be interpretable as A transforming gradually into B and then back again (as in figure 4.10e). Hence, besides the requirement that there should be evidence that sound A has been interrupted by B, there is also the requirement that B must not be interpreted as part of the same sound as A1 and A2. Otherwise, one sound should be heard.

This depends on whether a continuous change can or can not be *tracked* by the auditory system, and therefore also refers to measurement limitations of the auditory system. The interpretation of a gradual transformation is therefore only possible if the transformation is characterized by a slow and continuous change. This is comparable with the effects seen with alternated high and low frequency tones where an increased Δf and/or a decreased Δt led to the segregation of the tones in different streams. It can be interpreted as that the auditory system does not "believe" that a sound can be transformed into another one instantaneously without going through intermediate stages. This is one of the reasons why the noise burst B segregates from the pure tones A1 and A2 in separate streams. It is *very unlikely that a natural sound source could produce such an abrupt transformation in the production of two qualitatively very different sounds.*[47] Bregman proposes that such sudden changes in intensities triggers two "rules":

(1) The analysis of the current sound should be suspended and a new stream started, which occurs at the boundaries between A1 and B, and between B and A2, and
(2) The new stream should be connected to some former stream if possible, and if the connection is successful, the analysis of that continuing stream should be resumed, which leads to binding all the A's together and all the B's together.

The idea that sudden discontinuities induces the auditory system to begin a new analysis is supported by some findings that will be discussed in the following section concerning the presence of gaps between signals, and in section 4.7 in the context of the role of onsets.

However, the idea that the analysis of a stream, that already existed before the new stream started, is suspended, is not without problems. A better interpretation would be that *the auditory system always uses a broader temporal scope before definite decisions are made*, hereby allowing a consistent overall interpretation of the auditory scene reaching the level of perceptual awareness. In this case, the analysis is never really suspended, but a *temporary ambiguity is just ignored*. This will also be discussed in the following section regarding pitch perception in the presence of noise bursts, and in chapter 5 in the context of speech perception (schema-based processing). It will be shown that is seems that, as long as there is no counter-evidence present to hold the hypothesis that has been activated, this hypothesis not only stays activated, but also *continues to build up* the amount of activation.

Finally, it should be noted that it is hard to determine whether it is the masking potential of the signal or the fact that it includes hypothesis-confirming or *noncontradictive* stimulation that enables the restoration to occur. In the latter case, the hearing of the illusory continuity becomes just a particular case of hearing a continuing sound, since there is really nothing restored. The subject perceives the interrupted speech as continuous, because it is *the most likely interpretation given the evidence.*

[47] A better interpretation would therefore not be in terms of "believes" of the auditory systems, but in the auditory system *knowing* that such changes are not possible within a single natural sound source. This knowledge may be either innate, or result from learning based on prior experience with (many) natural sources. Furthermore, the excitation pattern of the BM to an abrupt broad-band noise source leads to clearly identifiable effects that correspond to the BMs responses to onsets in general, i.e., impulse response effects.

Many experimental findings support this conclusion. For instance, it has been reflected in articulatory interpolation (section 2.4.3), the experiments on the continuity illusion presented in this chapter, and the results on the phoneme restoration effect in the speech literature (sections 5.4.3.3 and 5.6.5).

4.6.5 Apparent continuity and pitch perception

In a recent investigation, Plack and White (2000) investigated what the influence of integration mechanisms in human pitch perception is on F0 discrimination in response to level discontinuities such as short periods of silence or periods of interruption by a 20 ms noise-burst. Usually, when information is combined over time, this leads to an increase in discrimination. So, there is an advantage in collecting information over a longer period of time. When analyzing the results of increased duration of a stimulus on the detection threshold of that stimulus, it has often been found that after 200 ms this threshold does not change anymore. Therefore, it has often been assumed that there exists a fixed and relatively long integration window (~ 200ms) on which the decision is based. The multiple-look model (Viemeister and Wakefield, 1991) assumes that the listener's decision in such tasks is based on multiple samples, or "looks" obtained through a short (3 to 5 ms) time window, which is a more flexible integration scheme. Other models with adjustable time windows or multiple time constants have also been suggested (Dai and Wright, 1995). In general, it seems that central decision processes mainly determine the outcome of such time-intensity dependencies (Dai et al., 2000).

White and Plack (1998) suggested that there are different mechanisms involved depending on whether the pitch has to be determined for (low pass filtered) *resolved* harmonics (up to 1875 Hz), or for (high-pass filtered) *unresolved* harmonics (5500-7500Hz). For the resolved harmonics, the improvement in discrimination was as would be expected from a multiple-looks model, i.e., an n times increase in duration, leads to an increase in d' [48] by a factor n. For unresolved harmonics, however, this increase was much greater, which has been interpreted as indicating the use of a long integration window to extract F0. However, if this were the case, presenting two 20-ms unresolved complex-tone bursts, separated by a gap, would lead to a same increase in d' as when a single 40 ms complex-tone burst was presented. It was demonstrated that this was not to the case for gap durations as short as 5 to 8 ms. Performance was much worse in the presence of a brief gap, and became comparable with predictions of the multiple-looks model, i.e., based on the combination of independent pitch samples. This was interpreted as indicating that the pitch mechanism for unresolved harmonics uses a long integration window that is *reset in response to level discontinuities*, thus leading to a loss of the benefit of long integration as a result of a brief break in the stimulus. A sensible strategy in the environment, where a discontinuity will often reflect the start of a separate auditory object that should be analyzed separately.

However, if a signal is interrupted by a stimulus such as a 20 ms noise burst that could act like a potential masker, then the auditory system normally assumes that, with no evidence to the contrary, the signal continues during the interruption. Therefore, if resetting depends on the detection of a level discontinuity, then placing a masker in the gap may prevent this resetting thereby *improving* F0 discrimination. Plack and White (2000) investigated this paradoxical prediction that adding a masker would improve F0 discrimination. It turned out that for the gapped conditions discrimination was significantly poorer compared to the continuous conditions, consistent with the result of White and Plack (1998).

[48] Within Signal Detection Theory (SDT), d' (pronounce as *d prime*) is used as a measure of discrimination, see appendix A.

However, when the noise was added, there was no significant difference. So, the presence of a noise did not appear to hinder pitch analysis in any way. Further manipulations confirmed the interpretation that the pitch mechanism *must* have been integrating across the gap in the noise condition. These experiments therefore suggest that *illusory continuity may be treated in the same way as real continuity as far as the pitch integration mechanism is concerned.*

4.7 The role of onsets and offsets

The onset and offset of a sound event is a very prominent, special case of the common fate principle regarding synchronous amplitude modulation (AM). All onsets consist of a periodic contribution and an aperiodic contribution. The periodic contribution is a weighted superposition of harmonic components, and may be absent. The aperiodic contribution can be characterized as an impulse. It therefore leads to a very short time-period in which all frequencies are simultaneously present on the BM. There is some ignorable phase-delay (between 2 - 40 ms) in the BM response to the low-frequency regions compared to the high-frequency components, probably due to the presence of the noisy aperiodic component masking the asynchrony of the onsets. The neurons connected to the BM are therefore synchronously activated. This causes them to remain in synchronized oscillation for a short period of time, which enables them to become perceptually more distinctive as an entity that can, as a result, more easily capture attention.

It is to be expected that onsets and offsets play a special role in auditory scene analysis, because it is exceedingly improbable that in the normal acoustic world unrelated sounds will start or end at exactly the same time. Lots of experiments indicate this special, independent effect of this type of synchrony. For example, in describing an earlier experiment in section 4.4, it was already mentioned that when the components B and C of a two-tone harmonic complex had different onsets and/or offsets, they tended to segregate more easily which made them more available to group with other signal components, i.e., to be captured by other signal components.

Also, the old-plus-new heuristic mentioned in section 4.6.1 is based on this improbability. When there is evidence that a certain signal continues at the next moment, but is then accompanied by other spectral components (that can not be interpreted as a gradual continuation of the preceding signal), these new spectral components are not interpreted as belonging to the old signal, but as signaling the presence of a new source. The auditory system is highly sensitive to differences in amplitude pattern for different partials and uses these differences to segregate them.[49] This way, partials can be captured out of a complex of (more than two) harmonics when one partial is preceding the complex. A later offset of one partial can also help it to segregate from the others. Such retroactive affects in simple perceptual phenomena have already been shown in the dependence of the continuity illusion on the part *after* the interruption.

[49] A phase difference between one partial and the rest of the partial also leads to that partial segregating from the others. It seems that the ear can instantaneously compare the phase of all the components and find subsets that are incongruent. It has been suggested that there is a mechanism in the peripheral auditory system that changes a phase difference into an amplitude difference (Kubovy & Jordan, 1979). Nevertheless, despite this apparent phase-sensitivity, it seems that phase matches are not necessarily for the purpose of deriving a pitch (but this absence of necessity holds for many sources of information), especially when there is a statistical independence in the phase changes, i.e., a fixed and repeating asynchrony is a much better clue for the presence of multiple sounds than an irregular asynchrony.

Furthermore, it explains the formation of residuals, not only for different spectral components, but also for sudden increases in intensity for components at particular frequencies. This is because of the *transparency of sound*, i.e., the fact that the mixture of signals we receive contains frequency components that are the result of multiple sources. In these cases, the AS not only has to ascribe a certain frequency *component* to a certain stream, but also has to "decide" how much intensity in a certain frequency region belongs to each stream that is present at the same time. Or, in particular, the *amount* of energy to leave behind to interpret an ongoing signal as undergoing *continuous* (and not abrupt, i.e., impossible) changes, and to allow the rest of the energy present at that frequency to contribute to another stream. Note that this poses some problems for the principle of exclusive allocation. Though this principle does have some validity for the visual scene where different objects can truly obscure each other, its validity is much less for the auditory scene, where different "objects" just add their intensities for shared frequency components. We will return to this issue later.

4.7.1 Voiced sounds

In case of the onset of *voiced* sounds there are at least two sources of information that the auditory system (AS) can use to group these sounds into the same stream stemming from a single source.

(1) One is the fact that the different harmonics are exhibiting the same AM over time, which is reflected in the temporal envelope.
(2) Secondly, the AS can use the beat rate between the neural outputs of the individual harmonics, i.e., their synchronous firing with the fundamental period of the speaking human voice, in addition to the detection of harmonically related places at the BM, corresponding to the same F0.

This may seem to imply a certain redundancy in information processing, but of course, as the human sensory system is capable of processing many features in parallel, it is likely that the AS is not concerned with such parsimonious arguments. In fact, it has already been shown in chapter 3 that there *are* many redundant mechanisms that complement each other. As no single mechanism can be trusted completely and different mechanisms may be more appropriate for dealing with different signal representations, this actually makes the AS so noise-robust.

An example in which both sources of information can be used, is the detection of the presence of the sound [z], which consists of an aperiodic component and a voiced sound. Here, the higher-frequency noise component is not harmonic and can therefore not be grouped with the voiced components based on a harmonic analyzer, but it is powered by the glottal vibration, which *is* periodic. Therefore, it does contain the same periodicity of AM as the voiced components of the [z] which can be detected by analyzing beats or local periodicities.

When there are multiple sound sources present at the same time that all have voiced components, the *differences* in onsets and offsets of the individual voicing events over time causes them to *segregate*. More generally, the *simultaneous* AM of the partials from the same F0, causes these partials to *fuse* into the same stream. In addition, as a result of differences in AM there are moments in time where one of the sources is dominating the spectrum while at other moments another source dominates the spectrum. Such momentarily glimpses or "peeks" of the spectrum, may be of particular importance for following one of the streams despite the fact that this particular stream at some points may be locally masked by the other streams.

So again, this points to the importance of:

- the *local* signal-to-noise ratio (SNR), i.e., not over the whole spectrum at a certain instant in time, but within smaller frequency channels;
- a *global analysis*, i.e., over a longer *temporal scope*, or the presence of some sort of *memory* (related to the old-plus-new heuristic);
- *using global properties* to help *tracking*, i.e., *attending*, to one particular stream.[50]

4.7.2 Psychophysical overshoot

An interesting phenomenon related to ASA and the role of onset asynchrony is that of *psychophysical overshoot*. In discussing the concept of masking, it has been mentioned that one signal component can mask another (simultaneously present) component. In psychoacoustical experiments, it has often been shown that in both cases the masking threshold depends on the duration of the background masking signal, i.e., the onset of the background masker, in relation to the onset of a superimposed target stimulus. As the onset-to-onset difference increases, there is a severe nonlinearity at the level of the auditory cortex in that the incremental response, i.e., the change in firing rate, to the target stimulus grows. This shows itself in a decrease of the masking potential of the background noise,[51] and thus creates a greater segregation between the effects of the two sounds. This phenomenon is known as psychophysical overshoot, and it has been shown that the amount of overshoot can be as much as 20 dB. Hence, in such cases the auditory system responds in a way that more qualitatively approaches superposition, because the different sound events are more distinguished as being *independent* auditory events (Smith et al., 2000).

Consistent with the experimental findings on masking described earlier, this illustrates that *the default of the AS is to integrate sensory evidence into one stream, unless there are cues to differentiate and therefore segregate different signal components*, such as in this case onset asynchrony. The AS is extremely sensitive to these segregation cues. Again, it also reflects the relative nature of the evaluation of evidence, i.e., the importance of taking a more global context into account instead of just looking at the particular frequency components that are present at one instant in time, or within a very small time frame.

4.7.3 Short-term adaptation

The role of the onset in the phenomenon just described might also be related to physiological mechanisms such as *short-term adaptation*, which is the combination of two effects:

(1) given a *constant* stimulus, the auditory nerve response decreases monotonically with increasing stimulus duration, i.e., it adapts and asymptotically reaches an equilibrium level, and

[50] Note that the principle of *common fate in AM* is therefore not a simple primitive *process*, though it is a simple *description* for summarizing the importance of many processes that (indirectly) influence one another based on *both* local and global analyses. This blurs the distinction between primitive and schema-based processes as defined by Bregman.

[51] This background "noise" can be either a simultaneously present tone nearby in frequency, or a noise band centred around the masked frequency component. The effect is therefore not attributable to global, qualitative differences between masker and target. Furthermore, in all of these kinds of experiments impulse-like onset effects are prevented as much as possible by slowly increasing the intensity (usually within a few ms.) of the stimuli instead of abrupt onsets, thereby corresponding more closely to the onset of natural sounds.

(2) *after* stimulus offset, a period of recovery in AN activity with a firing rate below spontaneous emission in quiet and reduced response to a new stimulus can be observed (Delgutte & Kiang, 1984b).

The latter refers to auditory fatigue, the former to auditory adaptation, or habituation. This short-term adaptation may contribute to the relative attenuation of an onset in the presence of other sounds.

Based on this kind of knowledge from psychoacoustics and physiology, the attenuation of *changes* such as shown in onsets has been incorporated into some ASR front end systems, by means of for example an adaptation loop (e.g., Tchorz & Kollmeier, 1999; Kollmeier & Derleth, 2000). That such prominent changes play an important role in auditory perception and therefore the usefulness of emphasizing these changes, seems very plausible in the light of the importance of signaling new events in the auditory environment, which has been described in section 4.1.1. Of course, this is also related to the fact that *the steady-state after an onset does not provide new information, whereas the dynamic signal properties do provide information.*

4.8 Global properties, selective attention and learning

An interesting effect that has been argued to *result* from ASA, which occasionally has been mentioned, is that *global properties* such as rhythm, temporal order, pitch, timbre, brightness, etc. only exist *within* streams. For example, when a sequence of tones, such as the alternating high/low sequences, or the galloping rhythm mentioned earlier, are segregated into separate auditory streams, the original rhythm is lost, and two streams with different rhythmic patterns result. Also, temporal judgments regarding temporal order between elements in different streams are much more difficult to make than within the same stream (e.g., Bregman and Rudnicky, 1975).

Therefore, Bregman argues, the sensory evidence stemming from a single source has to be integrated into a stream before such global properties can be "computed", which is done by primitive processes of auditory grouping. This is one of Bregman's arguments for claiming that *first* primitive processes of grouping are involved, and that schemas supplement the primitive processes. For instance, it has often been stated that the *cocktail-party effect* can be explained by following the pitch-contour of a speaker's voice, or by attending to the location in space, etc. It seems that once such global properties have been determined for the person that is speaking, we can *use* these to assist in subsequent grouping. Therefore, once "computed", properties such as pitch, can serve as the *basis* for *another level of grouping*, where selective attention is involved. This has also been referred to as *"tuning in"* to the characteristics of a particular speaker to allow the *tracking* of one speaker. Such global properties may therefore be considered as another source of information.

However, these global properties can not be computed *unless* the relevant signal components have been put into the same stream. This is because spectral components do not *have* timbre or pitch, for example, only sounds (which are the result of the perceptual fusion of the spectral components) have such global properties. Though this seems a strong argument, it actually seems to reflect how the same phenomena can be described at different levels of aggregation. Bregman assumes that global properties such as pitch need to be computed before they can form the basis of grouping.

An alternative interpretation would be that the harmonicity relation that exists between partials of the same fundamental leads to the perception of pitch due to the automatic grouping process that ensures the integration of these harmonics into one auditory entity.

This integration process of harmonically-related components does not *depend* on a computation of pitch, but the perceived pitch is just an emergent property of the grouping process that goes on continuously, i.e., it *is* the basis on which grouping occurs. Due to this perceptual fusion process, the sound as a whole becomes perceptually more salient, which allows it to capture attention. Under normal circumstances, this is strengthened by the simultaneous onset of the harmonic components (section 4.7). From this point, processes related to selective attention become involved, which usually means that the listener's mental state may be prepared in such a way as to enable the listener to segregate out a particular portion of a sound. In the case of pitch, this can be *described* as tracking the harmonics that correspond to the same pitch on the basis of pitch continuity. But, of course, the harmonics themselves also evolve continuously and simultaneously, thereby continuously providing the relevant physical cues to produce continuity in pitch. Never-theless, since the sound event has become the object of attention, it is made more easy to segregate it from concurrent sounds. Does this reflect different processes, or does it just describe the same process at different levels of description?

In the broadest sense, many of the manipulations that can be made to a portion of a complex sound (e.g., increasing the intensity of a single component, mistuning a component, modulating it differently, onset/offset asynchrony) might be regarded as a way of drawing attention to that particular portion of a sound. For instance, the ability to capture a harmonic out of a complex (an example of *analytic listening*) is argued to depend upon the liveliness of our recollections of the tones as heard separately (Seebeck, 1841). However, selective attention only involves the segregation of entities. It is not known whether it can also mediate the integration of an entity from a series of separated components (Hartmann, 1988).

4.8.1 Entity segregation on the basis of pitch

In this context, it is useful to realize that *stream segregation* refers to *a perceived association of separate entities into groups*, and therefore it is *not the same as entity segregation* (Hartmann, 1988). For instance, in the experiments on the streaming effect described in section 4.3, the auditory entities – the high and low tones – remained the same, whereas the resulting streams differed dependent on the experimental manipulation. Nevertheless, the segregation of entities and the segregation of streams share some features. For instance, rapid representation promotes the integration both of spectral components into an entity and of entities into a single temporal stream, and a slow representation favors both analytic listening and the segregation of entities into different temporal streams. However, in entity segregation a harmonic relationship among the components is a primary factor, whereas the usual stream segregation situation is tonotopically based with the (fundamental) frequency difference between (complex) tones playing the major role.

So, pitch plays a primary role in the segregation and integration of entities. Within models of pitch perception, the question of *integration* is usually addressed: How does the auditory system perceive a single pitch from a collection of harmonics with many different frequencies? Several pitch models haven been proposed that can be roughly divided into two classes: *spectral pitch* models in which tone complexes are neurally encoded as spectral patterns (due to cochlear analysis) and *temporal pitch* models where they are neurally encoded as temporal patterns (due to phase-locking to the individual components, or to the envelopes of their interactions).

Within spectral pitch models the perception of pitch is based on the lower harmonics that are resolved, where for fundamental frequencies between 100 and 400 Hz, the 3rd to 5th harmonics are the most important contributors, i.e., they are *spectrally dominant* (Ritsma, 1967). Since the pitch can also be perceived without the presence of the fundamental (Seebeck, 1841), for dichotically presented higher harmonics (e.g., 1600 and 1700 Hz leading to a

pitch of 100 Hz), and for different sets of harmonics of the same F0, this led to the idea that the perception of pitch is mediated more central by analyzing the pattern of resolved harmonics, and selecting the F0 that is most likely to have been part of that pattern.

For unresolved (higher) harmonics, the perception of pitch was initially explained by the residue theory of pitch perception as proposed by Schouten (1940). In this theory, the *beats* between adjacent harmonics lead to the formation of a *residue with a low-frequency envelope* of which the frequency is equal to the difference between harmonic frequencies (and therefore F0). However, when the harmonics are shifted by a constant frequency, a pitch shift is perceived, despite the fact that the envelope is unchanged (de Boer, 1956). This has led researchers to conclude that the *temporal fine structure in the residue* mediates the low pitch sensation: the time interval between waveform peaks near a maximum of the residue envelope codes for the pitch. A problem with this "peak picking" procedure is that it is very phase sensitive, since the time interval between envelope peaks depends sensitively upon the relative phases of the unresolved components. In normal listening conditions, these relative phases are highly unstable. Furthermore, this theory can not explain the existence of the spectral dominance region where harmonics are usually resolved.

Other models emphasize the temporal aspects by referring to the *temporal fine-structure of the neural pattern of oscillation*, where phase-locked neurons that are activated by harmonics of the same fundamental fire synchronously with one another according to the fundamental period F0. This information therefore directly reflects relevant pitch information (Schouten, 1974), and also corresponds to the perception of a low, residue pitch related to (high-frequency) non-resolved harmonics. In general, temporal algorithms utilize temporal response patterns in each auditory channel, and then combine information across channels to get the final pitch estimate. The nature of the cues and mechanisms that are employed correspond to, e.g., autocorrelations of the responses (Slaney and Lyon, 1993; Meddis and Hewit, 1991; de Cheveigné, 1998), or synchronization measures and oscillators (Langner and Schreiner). In a recent model, suggested by Shamma et al. (2000), both periodicity pitch and residue pitch *emerge* as the result of an integrated unitary mechanism based on a coincidence detection mechanism that instantaneously compares responses *across* channels.

4.8.2 The influence of learning

In general, within pattern recognition models, the underlying mechanisms involve the *recognition of harmonic patterns* in their influence they have on the neural excitation pattern. Such models are therefore influenced by learning as a result of very frequent exposure to harmonic complex sounds (e.g., Whitfield, 1970; Terhardt, 1970, 1974a; Goldstein, 1973; Schouten, 1974; Shamma et al., 2000; Wightman, 1973b). Such pattern recognition can be accomplished for example by means of a statistically-based *optimum processor* (Goldstein, 1973), where a *harmonic sieve* (or *spectral template*), acts as a *very* sensitive filter for those frequencies that are consistent with the same fundamental period. Only those frequency components can pass through the "sieve" (thereby also accounting for the segregation from other components). This requires that at least some of the subsequent harmonics have to be sufficiently resolvable by the pattern of excitation (either at the level of the BM or at the level of the AN fibers that are activated) and can therefore only deal with periodicity pitch.

The pattern transformation model, proposed by Wightman (1973b), on the other hand, constructs an autocorrelation function by using within-channel delay-lines, from a spectrum which is derived from auditory filters. It includes unresolved harmonics in the analysis along with the dominant resolved components. Also, the virtual pitch model (Terhardt, 1974a) is based upon a series of subharmonic cues of resolved pitches, and includes a role for analytic listening and the idea that the subharmonic transformation is learned. Finally, Shamma's model explains the formation of harmonic templates for periodicity pitch

perception as resulting from a process of learning so that cells in the AN and CN will be driven at many harmonically-related CFs. Hence, they will appear broadly tuned, but selective to a particular pitch value (Shamma, 2000).

Though there is still much controversy regarding the exact mechanisms responsible for the integration of harmonically-related components, and their segregation from inharmonic components, and whether or not it should be explained by a unitary mechanism, it seems obvious that spectral as well as temporal information can contribute to the integration and segregation process. The tonotopic organization at many levels of the auditory system, and the temporal response properties of the underlying neural processes allow for the emergence of a grouping of harmonically related components that can be described by the global property pitch. Though the basic mechanisms are innate, their fine-tuning and the possible emergence of harmonic "templates" are the result of a learning process. Given the fact that the ability to produce voiced sounds is general for natural sound sources, this can be easily learned early in development.

Likewise, it has been shown that the ability to group the spectral components under the envelope of a *familiar* sound tends to promote the integration of the spectral components under the envelope (McAdams, 1985). The existence of such an effect would clearly implicate processes at the highest level in the formation of auditory entities (Hartmann, 1988).

It is a well-known fact that the *effect of learning* leads to more efficient encoding, faster memory retrieval due to more efficient processing, a decreased involvement of attention and therefore an increased automaticity in performing the learned skill that is involved (e.g., Anderson, 1995). So, this makes it very arbitrary to *distinguish* processes or mechanisms involved in ASA in terms of innateness/automaticity versus learning/attention. Though the basic mechanisms that are needed are innate or *wired-in* (for example pitch detection and spatial localization mechanisms), this does not imply that perceptual learning (*fine-tuning*) has no influence on the proper functioning of these mechanisms. Of course, this also holds for more abstract source properties, such as a possible sensitivity to vocal tract dynamics. It is always difficult to distinguish between possibly *innate specific mechanisms* from our *general learning ability* or from *general processing mechanisms*, and it seems at least a bit arbitrary to base a theoretical framework on such distinctions. Therefore, the experimental findings that have been interpreted to provide evidence for primitive grouping principles operating before schema-based grouping actually indicate that the theoretical distinction between primitive and schema-based grouping is hard to maintain and probably irrelevant. They seem to reflect the same processing mechanisms that can be described at different levels of description.

4.9 Source properties and speech recognition

It has been demonstrated how simple processing principles seem to be at work while performing ASA. These principles relate to properties that generally hold for all sound sources in the environment. Interestingly as this may be, it has been shown with highly *un*natural stimuli (e.g., discrete tones, glides, repeated sequences), and in unnatural conditions, i.e., within an experimental context. The reason for using such stimuli is to force the stream integration and segregation, the fusion and separation of sensory evidence, that would be needed in the natural environment. Though this in itself does not pose any problem, it does raise the question if the derived principles hold when we use more natural stimuli. And, more interestingly in this context, whether it holds for processing *speech* in the presence of other sound sources.

Can we really extrapolate these (descriptively) simple and general principles to the complex, varying, uncontrollable natural environment we usually have to deal with? Does the speech signal truly exhibit these properties? Because if it does not, the generality of these principles could be questioned.

Furthermore, in how far is speech recognition affected by such primitive auditory grouping that has been argued to create the possible organizations, and vice versa: in how far are the groupings affected by incorporating specific knowledge regarding speech patterns? Are these general, simple principles really reflecting underlying general grouping processes and related mechanisms, or do they merely reflect convenient descriptions that are directly related to general source properties, without excluding the influence of more specific source knowledge during auditory grouping?

It is difficult to answer such questions, as working with real speech stimuli reflects the highly overlearned skill of speech perception, and therefore the heavy involvement of schemas. Nevertheless, a first approach that can be taken is to examine how manipulating the factors that influence primitive ASA, influence the perception of speech. Some of these will be discussed in following sections, but first some properties for which it is known that they hold for speech produced by one speaker will be summarized.

4.9.1 Continuity in pitch

One such property for voiced sounds is *pitch*. Continuity in the pitch contour is related to the fact that the pitch of one speaker changes gradually, and normally stays within a relatively small frequency region during the course of one speech utterance. Additionally, when two or more speakers are speaking at the same time, the different pitches (among other things, off course) differentiate these speakers, thereby causing them to segregate.

Furthermore, voiced speech is characterized by quasi-periodicity, and therefore each speaker is characterized by its own specific pattern of FM modulation (pitch fluctuations), which holds for all partials consistent with the same pitch, as they always change together. This makes that each instant in time the partials maintain their harmonic relationship. Even when two voices would have exactly the same pitch, the ear's apparent sensitivity to differences in the pattern of micro-modulation (or an extreme sensitivity of the pitch detection mechanism to exact harmonic relations or phase differences) makes it possible to segregate the two voices. Also, at a macroscopic level, the pitch modulation pattern, i.e., the intonation pattern, carries important information regarding the linguistic content.

4.9.2 Spatial continuity

Normally, the position of a speaker in space does not change, or at least changes relatively slowly. Therefore continuity in place, over time, can help in sequentially integrating sensory evidence. And, off course, when at a certain instant in time, energy at different frequency regions can be attributed to a source at one location in space, these frequency components are likely to stem from the same source, and therefore to become integrated as belonging to the same sound event.

In addition, when different frequency regions are related to different locations in space, it is likely that they do not come from the same source, which increases the probability that they will segregate.

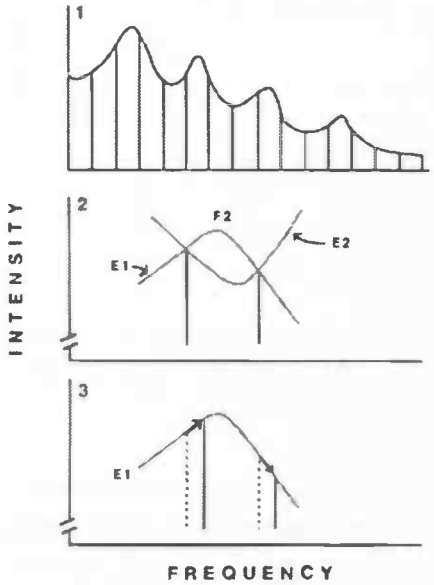


Figure 4.16 Illustration of "the tracing out of the spectral envelope" by a moving partial. (1) Spectrum. (2) Two possible spectral envelopes that fit the same harmonic intensities. (3) Evidence for the local slope of the envelope as the harmonics go up in frequency.

4.9.3 Formant trajectories

It has been described in chapter 2 that vowels can be defined by their formant *positions* (which is related to timbre), and that some consonants can be distinguished based on formant *transitions*. As the pattern of formants is the result of the filtering of the vocal tract, which can not change instantly, there is also a gradual change in the formant trajectory. Since formants are represented in the peaks of the *spectral* envelope, they refer to a more abstract, derived property of the physical signal.

It has been argued that they could be derived from the physical stimulus by a process known as *"tracing out the spectral envelope"* (McAdams and Rodet, 1988). When the pitch changes slightly, all harmonics change in a certain direction, so their amplitudes change also, corresponding to the attenuation characteristic resulting from the vocal tract transfer function. So, there is a covariation in the movement of harmonics. Information regarding such common fate changes can be used to estimate the positions of the formant peaks (figure 4.16). In case of a high pitch, i.e., a wider spacing between the individual harmonics, this peak estimation process is made more difficult.

Such spectral tracing is only relevant for the resolved harmonics of voiced speech, since for unresolved harmonics the spectral peaks are explicitly present in the cochlea representation of the physical signal. However, the auditory system can also "directly" represent such abstract properties such as formant peaks, which makes it possible to track the formant trajectories. It therefore does not need to depend solely on computations on *physical* signal components such as the amplitude changes of subsequent *individual* harmonics.

Physiological mechanisms that have been proposed can be based on the temporal pattern of neural discharges, such as is reflected in, for instance, the *Average Localized Synchronized Rate* (ALSR), *Post-Stimulus Time Histograms* (section 3.2), or the transformation of a phase-lock code to place codes in combination with lateral inhibition (section 3.9.2.1), on which the detection of patterns of FM sweeps could in turn be based. These are based on a temporal-place code. A rate-place code based on the average discharge rate of individual fibers can also represent formant information for unvoiced sounds and for unresolved harmonics (section 3.2.1).

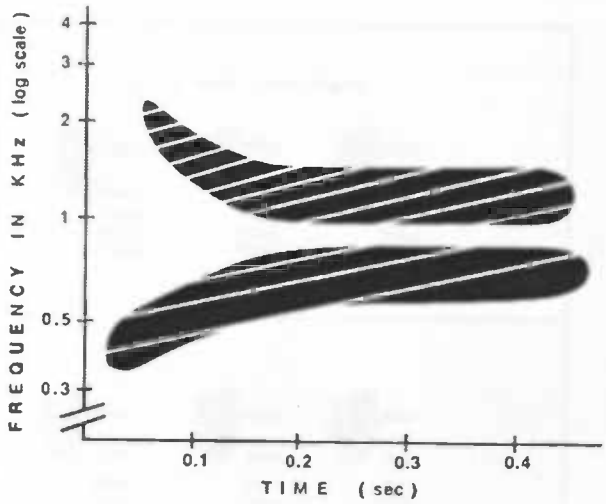


Figure 4.17 Schematic spectrogram of the word *yaw* spoken with rising pitch. The thin rising lines represent harmonics of a rising fundamental and the darkened regions represent formants.

This information is related to the dynamics of the human vocal tract, although FM sweeps are also common in the pattern of vocal inflections in the communication sounds produced by many animals. For the detection of FM patterns, processing still takes place at a subcortical level of processing and therefore reflects primitive processes of signal decomposition that are based on (several) signal transformations through an integration of time and frequency information.

Since the AS is able to use information regarding the formant patterns, it can "compensate" for the absence of pitch continuity for unvoiced speech by integrating sensory evidence where the same pattern of formant positions is maintained (or changes gradually) into the same stream. In the case of voiced speech, the formant pattern represents another source of information that *complements* the pitch information. Remember that the pitch and formants form *independent* sources of information that do not co-vary with one another. Different formants can move in different directions, and a rise in pitch, for example, can be accompanied by a dropping of one (or more) of the formants, and the onset and offset of different formants can be asynchronous (see figure 4.17).

The independence between pitch and formants has also been shown in an experiment by Marin (1987). It has been shown that a vowel, in a mixture of three vowels (/ee/, /ah/, /oh/), can be heard better if it is micro-modulated than if it is not. However, it does *not* matter whether the modulation is parallel or independent in the different vowels. So, if all vowels in the mixture are modulated with the same pattern of micromodulation, the identification of the vowels was as much facilitated as when they had different patterns of micromodulation. This lack of effect for modulated frequency coherence appears to suggest that the segregation of the vowels was not assisted by the *independence* of the modulation patterns (McAdams, 1984). But, when the vowels were steady, listeners did hear more than three pitches, probably because accidental harmonic relations in the mixed spectrum fooled the pitch-detection system into computing extra pitches.

Marin (1987) replicated these findings, and extended this research by asking whether it was the listener's sensitivity to tracing out of the spectral envelope, that strengthened the perception of modulated vowels as compared with unmodulated ones. Therefore, two forms of modulation where used, (1) a Trace condition in which the spectral envelope was fixed, and the amplitude changes in the moving harmonics could trace it out (figure 4.18a), and (2) a NoTrace condition in which the spectral envelope that defined a given vowel was not fixed, since it moved up and down in frequency with the harmonics, and the harmonics remained constant in intensity (figure 4.18b). These two forms of modulation strengthened the

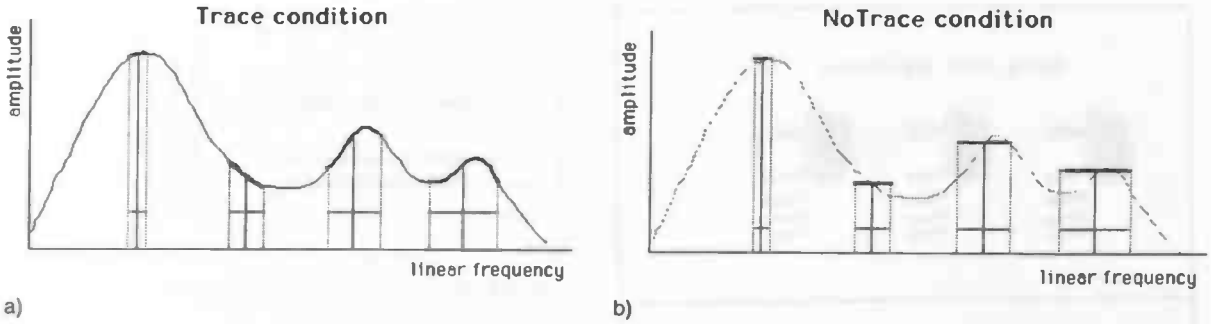


Figure 4.18 Illustration of two kinds of vowel synthesis. Amplitudes of frequency components are initially chosen according to a spectral envelope for each vowel. a) In the Trace condition, the amplitudes modulate with the frequency according to the spectral envelope. b) In the NoTrace condition, these amplitudes remain constant as the frequencies are modulated.

perception of modulated vowels equally well. This means that it is not necessarily the tracing of formants that makes modulated vowels clearer. But, as normally both modulations are present, these probably do strengthen one another in the perception of vowels, since they can provide independent sources of information.

It has been argued that the frequency-modulated vowels are more easily identified due to the greater perceptual integrity of the timbral percept compared to the steady-state ones. FM has been shown to increase perceptual integration for harmonically related sounds (McAdams, 1989; Darwin et al., 1994), though differences in FM cannot be used to segregate different sound sources (Gardner and Darwin, 1986; Gardner et al., 1989; Carlyon, 1991; Culling and Summerfield, 1995). This finding is one of many that illustrate the ineffec-tiveness of factors known to influence source segregation in affecting the recognition of speech sounds. The micro-modulation pattern might just be responsible for making the vowel formants *more speech-like*, and therefore induce a *phonetic analysis* related to *abstract knowledge* about known speech sounds and how they reflect vocal tract dynamics.

Such a phonetic analysis may therefore induce its own organizational principles, specific for speech sounds. Before discussing this possibility in more detail, there is another role that formant trajectories play in perceiving the speech signal as a coherent perceptual whole. Speech utterances comprise a variety of acoustic elements, periodic and non-periodic, continuous and discontinuous, exhibiting synchronous and asynchronous variation. According to the primitive organization principles mentioned thus far, it would be expected that these qualitatively different sounds would segregate from one another. Yet, the speech utterance is heard as a coherent whole. One factor that helps to perceive the speech as coherent is the presence of synchronous changes in the formant trajectories near the boundary between two different sounds.

An illustrative example is that, normally, when a mechanical click is superimposed on a sentence, listeners can not tell where in the sentence the click occurred (Ladefoged and Broadbent, 1960; Fodor and Bever, 1965). This suggests that the click and the speech are separated into *different* streams as it is consistent with the loss of between-stream temporal relations. However, when listening to, for example, a speaker of the language Xhosa (a click-language), it *is* possible to locate the position where the click occurs relative to the other speech sounds, even for someone whose native language does not contain such clicks. This probably relates to the fact that when a click is produced by speaking, it has a different acoustic relationship to the surrounding speech sounds than if it is simply superimposed onto the speech stream. These are reflected in a *synchronization of changes near the temporal boundary between the two sounds.*

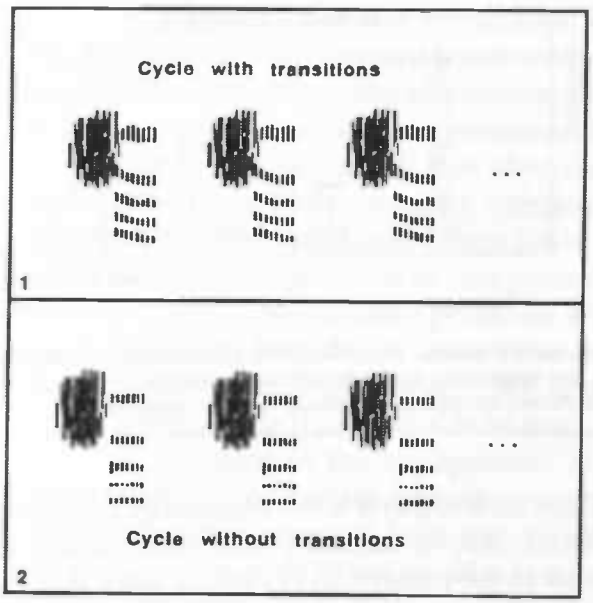


Figure 4.19 Schematic spectrogram of repeating cycles of the type used by Scott and Cole (1973). (1) Sequence of normal /sa/ syllable. (2) Sequence of transitionless /sa/ syllables.

Such changes near the boundary of two sounds also occur when for example a noisy /s/ is followed by the vowel /a/ making the syllable /sa/. The formant transition holds the stream consisting of these two qualitatively very different sounds together as one acoustic event. When the transition is deleted, the resulting sound when heard in isolation, sounds quite unintelligible.

In addition, when these transitionless syllables are played in repeated sequences, after a few cycles segregation occurs. A stream of /s/'s and a stream of /a/'s is formed. When the transitions are present, such segregation either does not occur, or only after a much greater number of repetitions (Scott & Cole, 1972). This suggests that the transitions in the formants play an important role for preserving coherence within a speech utterance. There are a number of factors that could be related to this:

- As can be seen in figure 4.19, the formants of the vowel all point in the direction of the noise burst energy. Therefore, a *backward, trajectory following extrapolation* might take place. Though retroactive effects in segregation have been shown, there is no evidence for a form of extrapolation based on the following of trajectories (see also section 4.6).
- Nevertheless, the transitions do prevent the occurrence of sudden discontinuities. Therefore, it may also be the case that their presence simply *preserves a form of continuity* because of the more gradual changes.
- Another possibility is that these results do not reflect the importance of the presence of the transition *within* the syllable, but merely reflect the effects of the used methodology, i.e., the *repeating cycles lead to an increased probability to group sounds over different syllables on the basis of similarity*, especially when the transition is absent.
- Finally, it might be that the *signal is being interpreted as a speech act*, since it is not possible for one vocal tract to produce such sudden acoustic changes, i.e., without a gradual transition from one sound to another. Hence, the transitionless syllable is not interpreted as speech which makes it sound so unintelligible, and which leads to the segregation on the basis of qualitative acoustic differences.

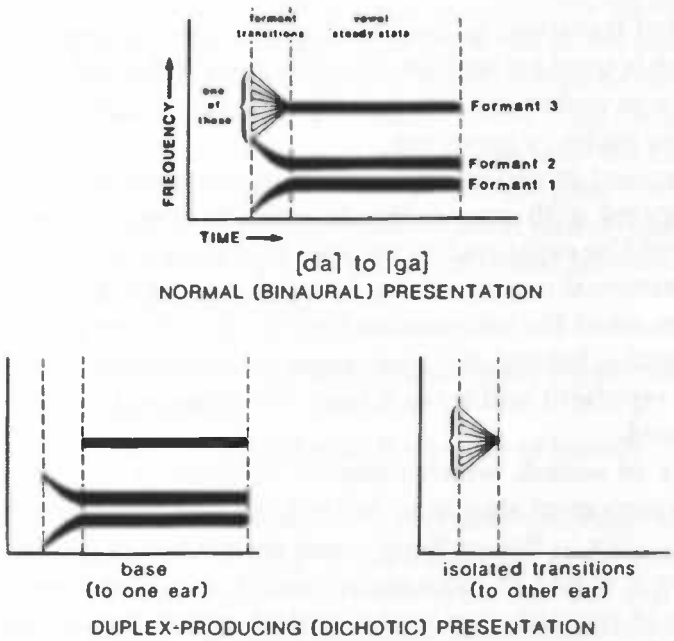


Figure 4.20 Schematic representation of a stimulus pattern used to generate duplex perception by Mann, Madden, Russell, and Liberman (1981). Top: the formant pattern used to represent "da", "ga", or sounds in between them, such as "ba". Lower left: the base. Lower right: the consonant-distinguishing formant transition. (From Liberman, 1982).

This latter possibility is especially interesting, because it suggests that the coherence of speech signals depends on their being perceived as resulting from a speech act. If it is perceived as speech, *additional* grouping principles seem to be involved. If it is not, grouping occurs on the basis of grouping principles related to the acoustic properties of the signal. Regardless of the interpretation, it is obvious that the formant positions and their trajectories can add an additional source of information, independent of pitch, for the integration of speech signals into a coherent whole.

4.10 Duplex perception of sound

Another interesting phenomenon is that of the *duplex perception of sound* (DPS). When for example, ideal formant patterns of the syllables /ba/ and /ga/ are compared to one another (see figure 4.20), it can be seen that they have the same formant pattern for the first two formants. They only differ in the transition of the third formant (F3). When this transition is isolated from the rest (the base), and then presented to the left ear, whereas the base is presented to the right ear, one syllable is heard. The identity of the syllable is determined by the isolated transition.

Therefore, although the information regarding the phonetic identity is presented to a different ear, it is integrated to form one phonetic percept that is heard at the ear where the base is presented. This happens regardless of a time difference between the offset of the transition and the onset of the rest of F3 in the base, and regardless of a frequency separation between the point where the transition ends, and the beginning of F3. So, *the identification of the syllable is not affected by dichotic presentation, time differences, and frequency separations*, factors that are know to influence source segregation. Instead, it appears to be affected by *phonetic coherence*.

In addition, however, the isolated transition is *also* heard as a "chirp" in the left ear. The fact that the third formant transition is used for both the identification of the syllable and for perceiving the chirp, therefore seems to violate the principle of exclusive allocation which is a very important principle in primitive auditory grouping.

For the simple stimuli that were used in earlier experiments, it was seen that the different cues for auditory grouping competed with one another in order to partition the data into separated streams. Many effects could be explained by the idea that one piece of sensory evidence could not attribute to two perceptual organizations. Nevertheless, it has also been mentioned that schema-based processes select the information they need, without partitioning the data, i.e., they do not leave a residue behind. Schemas seem to use sensory evidence that is consistent with the pattern they represent and leave it there for other analytic (primitive or schema-based) processes to be used.

In the light of the transparency of sound, which makes it difficult to decide how much intensity of a certain frequency component should be left behind, the principle of exclusive allocation seems to have less validity in the auditory scene than it has in the visual scene. However, even in the visual scene it would be more appropriate to consider the exclusive allocation principle more as a rule of contradiction in the context of building a consistent, coherent perceptual description of the environmental scene. The requirement then becomes one of the interpretation of local information as not violating conditions of consistency in the more global picture. There is a crucial difference between sensory evidence and perceptual descriptions. Sensory evidence is the raw input which the perceptual *interpretation* must account for. Perceptual descriptions are the accounts themselves. Sensory evidence can be multiplexed, shared out, or linked in different ways with different strengths. However, perceptual descriptions are constrained to be as definite and unequivocal as possible.

This distinction also explains the all-or-none inclusion of tones in streams, as tones are not sensory evidence, but a descriptions of a particular sound event. Therefore, the streams and the individual sounds within them are not created directly by primitive auditory scene analysis. Streams are descriptions, and primitive grouping merely establishes the links that bias the stream-building process in favor of including or excluding certain components of the sensory evidence. Thus, the principle of exclusive allocation does not describe the activity of primitive ASA as this process involves competition between groupings that do not have to be resolved yet.

Based on these considerations, Bregman (1990) explains the divergence of phonetic and auditory organization in DPS in terms of a two-component theory. Here, primitive processes can be considered as forming the hypotheses regarding possible groupings where the incorporated evidence has a more-or-less nature. The existence of two sets of links in DPS would be due to the fact that the sensory input from the base and transition stopped at different times, were separated in frequency, and/or occurred with different between-ear relations. This leads to a *bias* in favor of two distinct sources. Furthermore, the temporal proximity between the base and transition, and because neither input has more plausible auditory material to group with, leads to a cross-linkage between the two ears, favoring a single percept. So far, this is all the work of the primitive grouping processes.

Next, for building definite descriptions, schemas regarding distinct auditory sources in space and schemas regarding speech sounds, come into play. They decide in an all-or-nothing fashion about the definite properties. Connections are formed only by the primitive groupings or by *inclusion* in a schema, *not* by *exclusion* from a schema. For the speech recognition process, the subdivision suggested by the links would be overcome, whereas for the sources-in-space description the base and the transition would be heard as two separate sounds. *At this point*, the two descriptions would have to be composed with one another to form a more complete description. Because of the strong evidence for the chirp at the left ear as stemming from a single source, this will determine the hearing of that percept.

On the other hand, the speech schema takes all the information it needs to give a full description of the syllable. *Speech schemas do not seem to care about sources.* They represent abstract linguistic knowledge and seem to assume that if the information they represent is present in the signal, then it probably stems from the same source. As the right ear contains the most source information for the syllable, and the chirp at the left ear is probably too short to reliably estimate where it comes from, the speech syllable as a whole is interpreted as coming from the right, because normally speech sounds stem from a single source in space.

So, according to this account, there are two aspects of the descriptions created at the *same* level of processing:

(1) those concerned with the auditory properties of a source of sound with a spatial location, and
(2) those concerned with the sound as speech

The primitive level does not build descriptions at all, it only does preliminary linking of the data *prior* to the activity of more complex pattern-recognition processes, which usually has a strong influence on recognition processes but can be overcome by the latter if they have strong reasons for doing so. Normally, these two processes will converge on the same groupings of evidence.

Remember, however, that these primitive grouping principles have been derived from psychophysical studies of the formation of perceptual streams from spectrally stationnary and transient acoustic elements. Because of the acoustical complexity of speech, and because of the potential influence of the listener's linguistic knowledge, there is at least reason to contend the adequacy of the primitive ASA principles that have been discussed so far, to explain the perceptual integrity of speech signals (Julesz and Hirsh, 1972).

Furthermore, the way it is stated above, before recognizing speech, knowledge about the properties of speech sounds does not play an important role in the perceptual organization. This kind of knowledge would be supplementary to primitive auditory grouping which is based on domain-independent, local, acoustic cues only. Schemas therefore play a secondary role: they can confirm or repair the possible organizations, based on domain-specific knowledge. The emphasis placed on these primitive principles forecloses the possibility that *collateral* (or alternative) *resources* that are *no less fundamental* than auditory grouping principles promote organization of *some classes of sound sources* (Remez, 1994).

This emphasis on primitive processes of grouping may be unjustified, because there are many phenomena that show that the perceptual organization of speech depends on sensitivity to time-varying acoustic patterns *specific* to phonologically governed vocal sources of sound. Without the presence of any cue regarding the acoustic properties of the source, such patterns are capable to perform auditory scene analysis, selecting those parts of the signal that follow the speech-specific patterns, and leaving those that can not be interpreted as belonging to the speech sound, despite of the presence of acoustic cues supporting their groupings.

Therefore, some more phenomena will be described that show that phonetic knowledge can overcome the groupings suggested by primitive ASA, suggesting the independency of these two sources of information. Then, some phenomena will be addressed that show that phonetic analysis on itself is capable of performing ASA.

4.11 Split-formant research in speech

In an early research performed by Broadbent (1955), listeners were dichotically presented with *normal speech* that had been filtered differently at each ear. One ear received a signal that was high-pass filtered at 2 kHz, the other ear received the same signal low-pass filtered at 450 Hz. A large majority of the listeners reported hearing a single voice, despite the dichotic presentation. This could be attributed, however, to the synchrony of onsets of energy in the two channels when consonants were released, and by the common pattern of AM at the period of the fundamental in the two divided spectral regions.

Other research, using synthetic speech, has systematically investigated the importance of acoustic features in the grouping of formants. When for example the first and second formant pattern of a *two-formant synthetic replica* of a *sentence* were monochotically or dichotically presented, but differed in F0 (that followed a natural pitch contour), listeners reported hearing two voices saying the same utterance (Broadband and Ladefoged, 1957). When they did not differ in F0, and were dichotically presented, listeners report hearing a single voice at the side of the lower formant (Darwin, Howell and Brady, 1978). Having the same fundamental seemed to bind the formants together. However, when two generators were used, one for each formant, to generate the *same F0* pattern, fusion did not occur when they were presented to different ears. This was due to slight inaccuracies in the equipment which led the fundamentals to go in and out of phase randomly. This suggests a *great accuracy in timing* to induce the fusion of different spectral patterns, and just having harmonics that are related to the same fundamental is not enough to hear one voice.

Cutting (1976) further investigated the effects of F0 in binding formants together when presented to different ears. Even when the same pair of formants was sent to both ears, if the left and right ear differed in F0, two sounds were heard. When the difference between the fundamentals were as small as 2 Hz (i.e., 100 and 102 Hz), this already led to the perception of hearing two sounds. This *extreme sensitivity to small pitch deviations* has been shown in many psychophysical experiments.

When the first two formants of a *syllable* (such as /da/) differed in F0 and were presented separately to different ears, this again led to the perception of two sounds, but this did not prevent the listeners from being able to correctly recognize it. This finding has been repeatedly found. In every case in which a speech sound is created by combining formants with different F0s, despite hearing two voices, there was no tendency for this difference to inhibit the perception of the phoneme that resulted from the combination of the two formants (Darwin, 1981). Evidently, listeners combined the information from each resonance to form phonetic impressions, as they were *phonetically coherent*, despite the concurrent perception that each resonance issued from a different vocal source.

The only exception to this general pattern occurred with a particular pattern of four formants, F1-F4, where the combination of F1, F2 and F3, in isolation, leads to the perception of the syllable /roo/, whereas the combination of F1, F3, and F4 would lead to the perception of /lee/. The fact that F1 and F3 appear in both syllables, leads to a competition for the belongingness of these formants. Unless complete formants are used twice, the listener cannot hear both syllables. When the component formants of one of these syllables shared the same F0, this syllable was being heard. Nevertheless, this syllable could also be heard on a significant number of occasions without this F0 consistency.

Other experiments have also found that a common fundamental is not a prerequisite for the phonetic coherence of formant bands. Even a *tone* with an average frequency of no less than four octaves above the F0 can be integrated into a formant complex corresponding to a syllable, despite the spectral dissimilarity of tone and resonant components (Liberman, 1987).

Stimuli	Presentation	Result
Natural speech signal	- Dichotic: low- / high-pass filtered	- Single voice
Conclusion: No influence of dichotic presentation on perceptual fusion (Broadbent, 1955)		
Two-formant synthetic sentence replica	- Monochotic: F1+F2: - Diff. F0 - Same F0 - Same F0 (out-of-phase) - Binaural: F1+F2: - Diff. F0 - Dichotic: F1 / F2: - Diff. F0 - Same F0 - Same F0 (out-of-phase)	- Two voices, same utterance - Single voice - Single voice, weaker binding - Two voices, same utterance - Two voices, same utterance - Single voice at side of F1 - Two voices, same utterance
Conclusion: F0 determines perceptual fusion. Very sensitive to deviations in F0 in combination with dichotic presentation (Broadbent and Ladefoged, 1957).		
Two-formant syllables	- Dichotic : F1+F2 / F1+F2 - Diff. F0 - Dichotic: F1 / F2: - Diff. F0	- Two voices, single percept (even with small F0 diff.: 100-102 Hz) - Two voices, separate locations, single percept
Conclusion: F0 determines number of voices heard. Very sensitive to deviations in F0. Dichotic presentation with diff. F0 does not influence perceptual fusion. (Cutting, 1976).		
Synthesized syllables	- Binaural: multiple formants: - Diff. F0 - Binaural F1+F2+F3+F4: - Same F0 - F0 favoring one syllable: /lee/ vs /roo/ - Diff. F0	- Two voices, single percept - Formants with same F0 corresponding to possible syllable are fused - Phonetic fusion also possible for one syllable (N.B. Two syllables can not be simultaneously heard)
Conclusion: Phoneme recognition is possible without the assistance of primitive ASA processes, though performance can be improved by using information provided by primitive grouping. In general, schema's can look for the formant patterns that they need. F0 in different parts of the frequency spectrum has no effect. In case of competing phonetic interpretations in overlapping frequency regions, F0 helps to disambiguate (Darwin, 1981).		
Tone + synthetic formant complex	- Binaural	- Phonetic fusion
Conclusion: Phonetic fusion occurs with spectrally dissimilar formants with certain bandwidth and underlying F0 and tones. Only formant (peak) trajectories are important for phonetic fusion (Liberman, 1987).		
Sentence replicas + comodulating surplus resonance bands	- Binaural	- Sentence is heard, against background of chirps and bleats.
Conclusion: Phonetic fusion occurs only for signal components that correspond to a phonetic description that can be produced by one vocal tract. Coincidence of onsets or offsets, common F0 modulation, similarity in frequency excursion of surplus bands does not lead to phonetic fusion (Liberman and Studert-Kennedy, 1978).		

Table 4.2 Summary of results obtained with split-formant research.

This seems to suggest that, *in this case*, the only role for primitive scene analysis is to disambiguate between competing speech recognition schemas, which have already grouped the needed signal components from the mixture of sounds, without the assistance of primitive ASA grouping principles. It has therefore been stated that whatever mechanism is responsible for Gestalt-based grouping may be irrelevant to the disposition to integrate acoustic elements phonetically (Mattingly and Liberman, 1990). Schemas representing speech-specific knowledge take what they need to organize the auditory scene, and source information is only beneficial in disambiguating between competing organizations.

There is also evidence that spectral components can comodulate with one another without being fused, if the components do not correspond to a phonetic description. Liberman and Studdert-Kennedy (1978) reported a phenomenon in which a *synthetic sentence* was accompanied by arbitrary, *surplus resonance bands* that coincided temporally and often in cen-

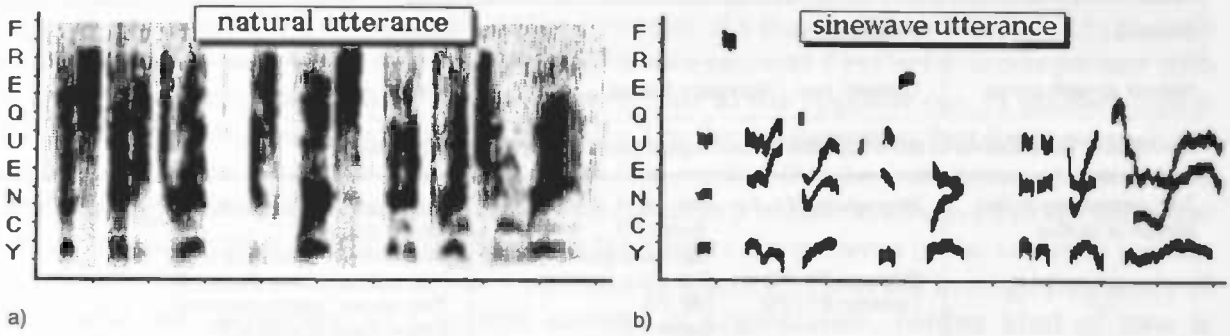


Figure 4.21 (a) Spectrogram of the natural utterance *"The steady drip is worse than a drenching rain"*, (b) Sine-wave replica of the same utterance.

ter frequency with the formants of the sentence. This was heard as a sentence against a background of chirps and bleats.

Neither coincidence in onset and offset, nor similarity of frequency excursions, nor modulation at a common fundamental were adequate to cause the listener to fuse the pattern into a single stream. The extraneous resonances neither contributed to the perceptual analysis of the phonetic properties of the sentence nor were they perceptually coherent with the vocal source. The results obtained within split-formant research described in this section are summarized in table 4.2.

4.12 Perceptual organization based on phonetic analysis

The problem with using (synthetic) speech stimuli in which there are both acoustic cues and spectro-temporal pattern cues is that the fact that speech is highly overlearned, may obscure any of the possibly independent contributions of both sources of information. Though speech *recognition* can occur in the presence of concurrent auditory groupings, this does not necessarily imply any *organizational* principles based on phonetic qualities. In a research performed by Remez et al. (1994) this question has been addressed by using *sine-wave replicas of speech*. They emphasize the perceptual effects of spectral variation, i.e., the *phonetic effects of spectro-temporal patterns*, derived from possible vocal effects.

It should be noted that, in this approach, phonetic effects are not interpreted as the detection of elementary acoustic particles that cue distinctions between consonants or vowels. Such particles are correlated with articulatory place, voicing, and manner cues (see section 2.2), and are exchanged for phonetic segments. The alternative view of Remez and others expresses that there does not seem to be a core of invariant acoustic cues (see also chapter 2); neither does the variability of the acoustic properties of speech signals implicate a normal set of acoustic tendencies around which cue variation occurs. Therefore, the acoustic cue is not an analytic entity drawn from a finite stock of vocalic effects, but the relevant description of acoustic variation pertains to the phonetic effects of spectro-temporal patterns (Remez, 1994).

Sine-wave speech consists of tone analogs for speech that include three or four time-varying sinusoids, each of which reproduces the center frequency and amplitude of a vocal resonance in a natural utterance (see figure 4.21). It therefore lacks the regular pulsing, aperiodicities, and broadband formant structure, characteristic of natural speech. Perceptual tests taken with this kind of stimuli reveal the effects of auditory variation independent of the momentary impression of vocal sound. The simple instruction to attend to a synthesized sentence is sufficient to permit an otherwise naïve listener to hear the phonetic structure of the natural utterance on which the sine-wave replica was modeled. But, even when phonetic impressions are evoked by the tone complexes, the timbre of natural voice is not.

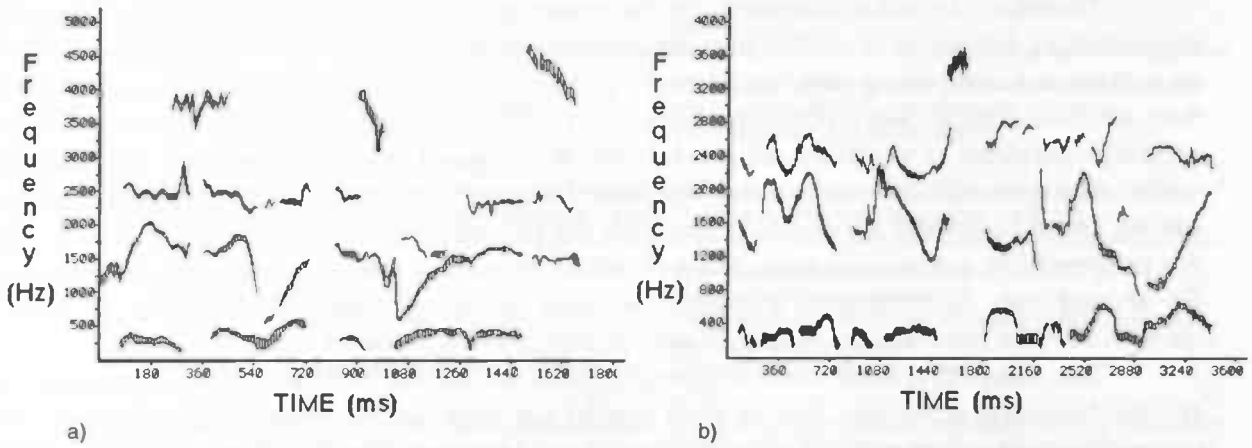


Figure 4.22 Examples of sine-wave speech utterances: (a) *"Please say what this word is"*, (b) *"The beauty of the view stunned the young boy"*. Note the absence of acoustic cues, and the inadequateness of cues for Gestalt-based auditory grouping to explain the perceived phonetic coherence of these kind of stimuli.

Nevertheless, it has often been shown that, as long as the perceiver treats the three or four simultaneous tones as if they were a complex acoustic effect of a phonological governed act, the frequency variation of the tones provides much of the phonetic information (e.g., Remez & Rubin, 1983). With these stimuli, the acoustic cues on which segmental distinctions are said to rely are absent, and the psychoacoustic properties of individual moments differ drastically from impressions of natural speech.

Remez (1994) argues that any convincing evidence that the *grouping* of several formants need not arise by comodulatory common fate (resulting from glotally modulated vocal resonances), diminishes the importance of this instance and exposes the Gestalt-based grouping principles as unimportant in the organization of speech signals. The perception of speech from such atypical auditory effects suggests that non-Gestalt resources describe speech abstractly and do not merely summarize typical auditory correlates of phonetic classes. Simple common fate is not required for grouping, not even when concurrent effects lack proximity, similarity, continuity and symmetry. It can be seen from figure 4.22 that, for example, there can be continuity in F1, despite intermittent discontinuities in higher formants (with gaps exceeding 75 ms), or vice versa. Further, there is a marked dissimilarity in the frequency trajectory of F1 and those of higher formants, and there is a lack of coincidence of large changes in resonance frequencies. However, to convey linguistic attributes, the simultaneously varying tones must cohere phonetically, the explanation of which warrants a *supplementary* principle.

While using sine-wave sentence replica's, subjects are asked to transcribe these strangely synthesized sentences. Often, the percentage of syllables that is correctly transcribed is taken as the performance score. Earlier research has shown that *phonetic perception requires the simultaneous presence of the first two formants, F1 and F2.* Neither the individual formants alone, nor the combination of F1/F3, or F2/F3 is sufficient to allow a transcription of the sentences (Remez et al., 1981).

It has been shown that babble masks (i.e., speechlike components) can impede more effective on grouping than for example broad-band noise masks (e.g., Carhart et al., 1975). Furthermore, the interference effect of babble seems to correlate with *distinctiveness*. For example, an extraneous channel composed of 3 voices interferes more with the perception of the target speech than babble composed of 16 or more voices. These results might be interpreted as reflecting the influence of masking, phonetic interference in immediate memory, or as being the result of competition in perceptual organization.

Therefore, in order to investigate the presence of a time-varying principle of phonetic organization, Remez et al. (1994) presented F2 to one ear, and F1+F3+F4, accompanied with an extraneous competing tone for F2, were presented to the opposite ear. In another condition, all tone analogs were presented binaurally. On the presumption that listeners seek the acoustic correlates of vocal sound production, it is argued that an extraneous sinusoid that varied in a speechlike manner within the same frequency region in which F2 in the other ear varied, should compete for organization with the proper components of the sentence replica. An unspeechlike extraneous tone, like a constant-frequency tone at the average frequency of F2, should not. According to a general account of organization, neither kind of tone is enough like the constituents in a replicated sentence to be grouped with.

The competing tone was a temporally reversed analog of F2 that has a close match in spectro-temporal attributes, but by itself would not likely evoke phonetic impressions, and was unlikely to cohere with F1+F3+F4. The reason for presenting the competitor tone dichotically was to exclude the possibility of the two different F2 analogs masking one another. An earlier experiment had already verified that the absence of the "correct" F2 in the ear containing F1+F3+F4 was not interpreted as being the result of masking. Namely, it was shown that the presence of a foil in the other ear, that could have masked F2, led to the same poor transcription performance as when F2 was just absent, i.e., when only F1+F3+F4 were binaurally presented. Hence, there was no evidence for phonemic restoration due to mas-king, whereas the presence of the correct F2 analog in this ear did lead to an obvious increase in transcription performance. This suggested that dichotic fusion had occurred, but only for the phonetically coherent stimuli.

In the dichotic competition condition, it was found that there was only a significant competition effect for the speechlike competitor, suggesting that the effect depends on the pattern of variation of the competing tone. The constant-frequency tone led to the same transcription performance as when no competitor was present. Remez et al. argued that the dichotic presentation of the F2 analog and its competitor permitted the segregation of the unspeechlike tone from F1+F3+F4 (sharing spatial locus, as they were presented to the same ear), and allowed the incorporation of F2 of the other ear. This did not seem to be possible, or at least much more difficult, for the speechlike competitor for F2 that was presented at the same ear as F1+F3+F4. In the binaural competition condition, either kind of competition led to a severe decrease in performance, which is likely the result of peripheral masking (both around F2, and as resulting from the upward spread of masking). In the dichotic condition, the F2 analog was freed from such masking effects such that it competed for organization with the speechlike competitor.

These results were therefore interpreted as follows. The grouping on the basis of speechlike variation apparently holds fast until a stage in perceptual analysis at which the phonetic information contributed by the actual F2 analog (at a separate spatial location) was less accessible. Furthermore, though the grouping by similarity in location is anticipated in the Gestalt approach, the pattern-continguent suspension of this rule is an effect of the reliance on a principle of grouping sensitive to the acoustic products of vocalization. Thus, the Gestalt-approach has problems in interpreting the dichotic results, for, despite the similarity in location, grouping does occur with no competitor and with an unspeechlike competitor across the ears, but not with the speechlike competitor.

The principle of similarity in frequency variation can not explain such grouping of tones to form sentences, in both binaural and dichotic conditions. In addition, the role of schemas as supplying information about typical acoustical properties of speech also does not seem to hold here. Sine-wave replicas fail to qualify as typical, and the rapidly fading auditory trace (estimated around 200 ms, e.g., Pisoni, 1973) of sine-wave replicas is not sufficiently durable to sustain a successful schematically driven inspection. Even when schemas are seen as representing very abstract phonetic knowledge corresponding to known groups of formants, the results indicate that such analysis has to take place very early in processing.

Presentation	Stimuli	Result
- Binaural - Dichotic	- F1+F3+F4 - F1+F3+F4 / F2 foil (reversed speech) - F1+F3+F4 / F2 correct	- Same poor performance for the binaural condition without F2 and the dichotic condition with F2 foil - Dichotic fusion only for correct F2
Conclusion: Only the correct F2 led to dichotic fusion, and therefore to an obvious increase in phonetic transcription performance. The F2 foil was not interpreted as possibly masking the correct F2, since no phoneme restoration occurred.		
- Dichotic	- F1+F3+F4 - F1+F3+F4 / F2 correct - F1+F3+F4+F2 const. freq / F2 correct - F1+F3+F4+F2 speech competitor / F2 correct	- Poor transcription performance - Dichotic fusion with correct F2 - Dichotic fusion with correct F2, same performance as without F2 tone-competitor - Competition, no dichotic fusion
Conclusion: Only competition in grouping with speechlike competitor such that dichotic fusion is prevented. With no competitor or a nonspeechlike competitor dichotic fusion with the correct F2 occurred based on phonetic coherence despite different spatial location. Competition effect in phonetic organization is dependent on the pattern of variation of the competing tone.		
- Binaural	- F1+F3+F4+F2 const. freq. + F2 correct - F1+F3+F4+F2 speech competitor + F2 correct - F1+F3+F4+F2 correct	- Same poor performance. Both types of competition prevent phonetic fusion due to masking - Phonetic percept
Conclusion: No competition in grouping occurs, probably due to masking, such that temporal pattern of correct F2 can not be accurately followed.		

Table 4.3 Summary of results obtained with synthetic sine wave speech (Remez, 1994).

It does not *depend* on any form of primitive auditory grouping parsing the sensory array into *sources* of sound. It is therefore argued that phonetic organization diverges from ASA early in perception and proceeds independently.

It seems that to group the auditory effects of speech production, the listener *rapidly* detects acoustic variation *consistent with an articulating vocal tract*, and cannot project from a proximal stimulation to distal vocal objects by resorting to the impoverished similarity principles or typicality schemas offered in ASA. The coherence among the components of a sine-wave sentence replica is established because the listener detects their plausible origin in a vocal source. Therefore, the grouping based on such abstract knowledge regarding vocal tract dynamics (without relying on particular recognition schemas), is as fundamental as, and independent of, the grouping related to the representation of sound sources based on acoustic properties. This independence is strengthened by the examples mentioned in this section and the preceding two sections. DPS can in this sense be interpreted as resulting from scene analysis and phonetic organization that occur independently when the components of an acoustic display give rise *simultaneously* to impressions of a single phonetic and multiple auditory sources (Mattingly and Liberman, 1990).

The results of the research presented in this section indicate the importance of testing the general auditory account for ASA using stimuli that represent the spectro-temporal variation that is characteristic of speech. Furthermore, the simple mechanical changes in the vocal source can create multiple acoustic effects for the perceiver (e.g., Rubin, Baer, and Mermelstein, 1981). Without the disposition to treat the signal as the product of a complex object, one with multiple sources of excitation and multiple reshapable resonant cavities (see section 2.1), and to ignore the dissimilarities and discontinuities among acoustic constituents, the speech signal would fracture into the bits that the general account warrants.

Finally, it should be noted that the fact that in the binaural condition the phonetic fusion with the correct F2 does not occur, as a result of masking by its simultaneously present competitors, indicates that sine wave speech is easily deteriorated by the presence of other signals. However, it has never been claimed by Remez et al. that sine-wave speech is comparable to the acoustic properties of normal speech, but only that its spectral-temporal charac-

teristics correspond to those of formant peaks. Since the locations of formant peaks and the pattern of formant trajectories define the phonetic identity (in context), it is generally believed that these play an essential role in the recognition of speech. Furthermore, in normal speech, the formants form the most noise-robust source of information. This is due to the fact that they result from the resonant properties of the vocal tract, and therefore their underlying spectral components are amplified, i.e., their local signal-to-noise ratio is raised.

Though the formants are not directly present in the physical signal, the auditory system has been shown to be able to represent formants based on an integration of time and frequency information that can be encoded by the temporal response characteristics of (groups of) neurons in the AS. Therefore, the information that is represented by formant peaks is available early in auditory processing. The analysis of the spectro-temporal pattern that characterizes the combination of different formants could therefore also take place early in processing.

Many perceptual findings indicate that the human auditory system can accurately track the vocal tract dynamics that characterizes human speech. Remez (1994) argues that the auditory system's sensitivity to how formant patterns can change together, is indicated by the fact that in the dichotic condition, perceptual fusion only takes place for the *phonetically correct* F2 pattern and not for other present signal components (even if they also have the temporal characteristics of speech and if their spatial location favors their integration over that with the correct F2 presented to the other ear). However, for the organization of these formant patterns, the primitive grouping principles that have been derived for primitive ASA can not be applied. Synchronous on- and offsets, parallel frequency changes, correlated AM modulations, and pitch information do not characterize the pattern of the different formant trajectories that stem from a single source. The only Gestalt-based cue that could apply is that of spatial location, but the results with the dichotic presentations indicated that this cue can be overcome in favor of a coherent phonetic percept.

Therefore, a plausible interpretation of the grouping of the sine wave speech sentences is that they are phonetically coherent, because they are *consistent with how the resonant properties of the vocal tract can continuously change together*. This reflects a sensitivity of the AS to very abstract source properties. However, this source property is no less fundamental than the acoustic properties that characterize normal speech. The multiple sources of excitation in speech production make that the speech signal consists of qualitatively very different sounds, which would according to primitive grouping principles easily lead to the loss of perceptual coherence. But, the constantly present and continuously changing filtering properties of the vocal tract ensure that there remains phonetic and therefore also, perceptual coherence.

4.12.1 The relation between primitive ASA and speech recognition

The findings regarding duplex perception, split-formant research, and the results from studies using sine-wave speech replica, have been interpreted by some researchers (e.g., Mattingly & Liberman, 1990; Remez & Rubin, 1983; Remez et al., 1981, 1994) as showing that ordinary grouping principles of auditory scene analysis do not apply to speech perception. Liberman, among others, argues that *speech is special* in that speech perception has its own rules for the integration of sounds, perhaps due to its access to knowledge about how speech is articulated. Arguably, the representation of speech that is needed for speech recognition should abstract away from the particular source characteristics of a certain speaker in order to be generally applicable (without completely eliminating this kind of information, of course).

When looking at the information-bearing elements in speech, e.g., formant locations, formant transitions, and noise bursts (see also section 3.9.2), it seems that the spectro-tempo-

ral pattern of these components relative to one another can not be described by Gestalt-like principles. Also, the global qualities of different speech sounds differ drastically. Nevertheless, speech events are perceived as a coherent whole.

Furthermore, despite the fact that the formants are not physically present in the signal, it has been found that the auditory system is capable of representing the spectral envelope. Therefore, formants are neurally encoded. Also, formant transitions corresponding to FM sweeps can be processed early within the auditory pathway. In the previous chapter (section 3.6), it was mentioned that two parallel functional pathways exist in the subcortical, central auditory system, one characterizing acoustic events by accurately encoding physical stimulus properties and related to source localization, the other displaying associative learning and important for labeling stimuli with perceptual qualities. As has been mentioned these have often been characterized as encoding respectively the "where" and "what" of acoustic events. This is consistent with the findings that indicate that in order to activate the schemas needed for speech recognition, it is not necessary that the underlying source characteristics are consistent with one another. Of course, for a complete representation of acoustic events present in the auditory scene, these sources of information are eventually integrated. The integration of this knowledge allows us to identify speakers, and to use the speaker characteristics to disambiguate between different interpretations of the same speech signal.

Notice, that *the acoustic cues that are important for, for instance, pitch perception and spatial localization, are important for information-processing in the brain in general*, as is reflected in the ability of neurons to phase-lock to the stimulus based on periodicity information, and to synchronize with one another to allow coincidence detection mechanisms to operate. These (and other) temporal response characteristics are general for a large groups of neurons and allow for the noise-robust transmission of information such that primitive signal components can be estimated. This occurs in specialized regions in the brain The *grouping* of these primitives can be aided by the presence of the same underlying pitch (by virtue of their synchronization, the increased perceptual salience, and selective attention processes). However, this is not sufficient nor necessary for the perceptual coherence of speech.

Another interpretation regarding the role of primitive ASA is that the strong sequential expectation, built up through experience with the language, overcomes any segregative tendencies. Either of these views makes bottom-up ASA almost irrelevant in language perception. However, it has been shown that differences in pitch and spatial location actually *do* make it easier to segregate two voices perceptually and to track the speech of only one of them, which are facts that are consistent with ASA. So, there is no reason to exclude primitive ASA in a normal acoustic environment. Especially for voiced sounds, where the harmonicity principle, and the common onset and comodulation of harmonics, directly reflect source properties, and provide a useful source of information.

Then, what could be the status of ASA based on primitive grouping principles? Remember that Bregman considered bottom-up ASA as establishing the links that vary in strength. If these are not too tight, recognition processes, looking for specific sets of components, can select the components despite their links to other ones. It is at least obvious that segregation can not be viewed as a *complete* packaging of evidence *preliminary* to recognition. Approaches to use bottom-up ASA as a front end to recognition processes in ASR without redesigning the recognition models to accept linked data as input rather than signals are therefore likely to fail. However, even when bottom-up ASA is viewed as only a hypothesis-generating process regarding possible groupings of data, it is very unlikely that it provides the only source of information for recognition schemas, which is exemplified by the fact that phonetic recognition occurs despite the absence of correct grouping hypotheses.

Another possibility is to consider the bottom-up and top-down processes as operating together, so that the description-forming process tries to satisfy the constraints resulting from evidence regarding different sources in space, and from evidence regarding speech utterances produced by one vocal tract, *at the same time*. This also implies a weaker role for

bottom-up grouping as it has been suggested by Bregman (1990). Actually, Bregman (1998) acknowledges that this may well be the case.

More recent developments in computational ASA therefore see ASA as such a combination of bottom-up and top-down processes. For example, in a so-called prediction-driven ASA model proposed by Ellis (1996), hypotheses that are generated bottom-up on the basis of *both* periodicity information and the time-frequency envelope, activate schemas that predict newly incoming information. These top-down generated predictions are being reconciled with the new bottom-up generated input. If the two are consistent, the activated schemas remain active, if they are inconsistent, this leads to a modification of the active hypotheses. This top-down flow of information probably leads to the same organizations as when learned patterns between (abstract) signal components (related to the regularities that result form vocal tract dynamics) are partly recognized early in processing such that they can constrain the schemas that are activated immediately in combination with a consistency of the underlying acoustic source characteristics. If the schemas become and stay active only when the bottom-up input is not inconsistent with them, there is no need for reconciliation processes at the level where new input comes in. Functionally, however, these differences in processing probably lead to the same system behavior, but the underlying mechanisms are different in character. The issue of how to characterize the information flow between bottom-up and top-down sources of information will also be discussed in chapter 5 (section 5.6).

To conclude, it is clear that some sort of scene analysis is to be done in order to make sense of the information in the environment. It is also clear that processes related to information stemming from different sources, and processes related to processing speech-specific knowledge both add relevant information in order to achieve this task. If these sources of information are processed in parallel, the encoding of these different sources of information does not seem to *depend* on each other, though they do need to be integrated eventually in order to achieve a coherent interpretation of the environmental scene.

4.13 Summary

Our perceptual faculties evolved as a means of allowing us to construct a useful representation of reality. Perception is functional and ecological providing us with the "what", "when", and "where" of the events around us. According to Bregman (1990), the primary task of the auditory system is to arrange the mixture of frequency components stemming from multiple sound sources into meaningful perceptual organizations that correspond to various real-world activities.

Some sounds mark the occurrence of unique events. But the world of sound is not merely a succession of momentary incidents. Even a sequence of discrete sounds are often caused by one on-going coherent activity. Most sounds have a history. The mental images we form of such "lines of sound" are called *auditory streams*, and the study of the behavior of such images is the study of *auditory streaming*. Since it is argued that the recognition of events depends upon the proper assignment of auditory properties to different sound sources, auditory streaming is fundamental to the process of scene analysis. This has led Bregman to propose that the auditory system accomplishes the task of auditory scene analysis by a primitive, automatic process in which grouping hypotheses are formed containing the information stemming from one source. These in turn are supplemented by domain-specific schema's that represent knowledge of familiar sounds based on prior experience.

Auditory streaming entails two complementary domains of study. How individual sounds cohere to form a sense of continuation is the subject of *stream fusion*, or sequential integration (resulting in *auditory streams*). Since more than one source can sound concurrently, a second domain of study is how concurrent activities retain their independent identities - the subject of *stream segregation* (or perhaps more accurately, *entity segregation*). In general, individual sounds tend to coalesce into a single percept in proportion to the physical correlations shared by the parts that reflect properties of the source. Stream-determining factors include: timbre (spectral shape), fundamental frequency (pitch) proximity (or continuity), temporal proximity, harmonicity, intensity, and spatial origin. In addition, when sounds evolve with respect to time, it is possible for them to share similarities by virtue of evolving in the same way. In Gestalt psychology, this perceptual co-evolution of parts is known as the *principle of common fate*. Bregman has pointed out that the formation of an auditory stream is governed largely by this principle. This is reflected in, for instance, the role of the common fate in AM principle. Especially, the simultaneous onset of certain signal components plays an important role, since it signals the presence of a new auditory event stemming from an other source, or a new auditory entity within an auditory event stemming from the same source. It is also related to phenomena such as psychophysical overshoot, and physiological mechanism such as short-term adaptation, and the synchronous firing of neurons. This makes them perceptually more salient, and therefore allows for these components to "draw attention".

Regarding common fate in FM, the observed findings seems to be more appropriately described by the principle of harmonicity, corresponding to the global property pitch, which plays a role for voiced sounds to integrate and/or segregate into separate auditory entities. Furthermore, proximity in pitch (F0) seems to be the determinant factor for sequentially grouping different auditory entities into auditory streams, at least for very short time-intervals.

The findings on the release of masking and the continuity illusion provide strong indications that the auditory system is generally gathering data within a more global temporal scope while it is extracting features (information) signaled by multiple acoustic cues. Temporary ambiguities resulting from masking phenomena are ignored as long as the neural acti-

vation pattern resulting from the acoustic stimulus is not inconsistent with the overall interpretation of the perceptual scene.

In general, it seems that important aspects in the grouping of signal components are the local SNR within frequency-channels, a global scope in time as well as across frequency-channels (such that context always matters), and the influence of attention. It has been argued that the distinction between primitive and schema-based grouping processes on the proposed dimensions are hard to maintain and probably irrelevant.

Besides the source characteristics that can be directly computed from the physical signal (such as pitch and location in space), other source characteristics are also important. This is reflected in findings such as the duplex perception of sound, and results obtained within split-formant research, and by using synthetic speech sentence replicas. The auditory system's ability to represent abstract signal properties such as formants and formant transitions subcortically is consistent with the interpretation that the spectrotemporal variation of these formants relative to one another allows for the grouping of these components early in processing. This ability is argued to result from an exquisite sensitivity of the auditory system to abstract patterns of regularities resulting from vocal tract dynamics, which is a characteristic of the human speech production source that is no less fundamental than the underlying acoustics characterizing voices.

In the normal auditory scene, both sources of information are available. Interestingly, the combination of these sources allows the auditory system to be very noise robust. For instance:

- the presence of voicing (i.e., pitch) ensures the perceptual integrity of individual sounds and makes them perceptually more salient so that they can become the object of attention; this is possibly related to the phase-locking capability of neurons, the ability to compare the temporal characteristics derived from the signal within-channels through performing autocorrelations, and across-channel coincidence detection mechanisms, and the possible existence of (learned) spectral templates
- dynamic aspects of the signal are emphasized, such that synchronous onsets of different signal components are capable of drawing attention to these components such that they can be more easily segregated; this is related to, for instance, psychophysical overshoot, and short-term adaptation
- the regions in which formants occur are amplified due to the vocal tract filtering characteristics, thereby ensuring a higher *local* signal-to-noise ratio

Therefore, despite the fact that the experimental findings presented in this chapter have been obtained by using atypical stimuli in unnatural settings, it seems that they nicely complement each other, and converge onto the knowledge derived from other research areas as described in previous chapters. The functional characterization of the mechanisms needed to perform ASA as originally proposed by Bregman, seems to be misleading and inadequate when referring to the underlying mechanisms. By taking the explanation of the experimental findings beyond a functional level of description, it is possible to give a more accurate explanation (or characterization) of the processes and mechanisms that are involved while performing ASA.

Chapter 5 *The Speech Processing System*

This chapter will focus on the *processing architecture* and *functional organization* of the speech processing system, where the emphasis will lie on knowledge derived from *psycholinguistics*. This interdisciplinary field is concerned with psychological aspects of language studies (see for example, Kess, 1992). One of the goals in psycholinguistic research on speech recognition is to understand the process of mapping acoustic input onto lexical representations. Some important questions that have been addressed are:

- What are the *units of representation* formed during this process?
- What is the *influence of contextual information* on the processing of perceptual input?
- What is the *direction of information flow* between different levels of representations?
- Are *commitments to lexical hypotheses* made *immediately* or are multiple hypotheses being preserved with *delayed* commitment to a final single lexical hypothesis?
- How is a continuous speech stream being *segmented* into a discrete sequence of words?

Within psycholinguistics, it is assumed that the detection of relevant acoustic-phonetic features has already taken place. Questions that are related to how these units are extracted from the mixture of sounds present in the auditory scene, and how these are grouped together are not addressed. These are considered to be research topics for other research disciplines. As a consequence, it is implicitly assumed that the problem of auditory scene analysis is not relevant for understanding the process of spoken word recognition.

An inherent "problem" with the nature of the questions mentioned earlier, is that these can only be addressed *indirectly* by testing subjects within a certain experimental context. Obtained results are then interpreted within some theoretical framework, and therefore conclusions are always "colored" by the assumptions underlying this framework. Within psycholinguistics, several experimental paradigms have been developed. Though conclusions may only be valid in the context of these paradigms, when different paradigms converge onto the same experimental findings, the validity of the conclusions is strengthe-ned. In the discussion of the experimental evidence, the experimental paradigms and their underlying assumptions will be described (where needed) to provide the means to put the evidence in the right perspective.

A general problem with perceptual studies, is that the behavioral results often reflect the *outcome* of perceptual processing, i.e., the perceptual interpretation of which we become aware (see also chapter 4). Nevertheless, attempts are made to investigate the *time course* of the underlying perceptual processing that takes place. Further insights can be gained, for instance, by analyzing the nature of the responses in more detail by means of reaction-time (RT) analyses. Within psycholinguistic modeling, different kinds of descriptive models exist: models that do not describe the time course of speech processing, and models that do make explicit assumptions on how different processing mechanisms are related to one another during processing, and how they influence the perceptual outcome.

However, these latter models can not be verified or distinguished without the ability to test the capability of a computational implementation of the model to account for the experimental data. Therefore, there are several computational models in which the assump-

tions about the underlying processing architecture and functional organization are formalized and implemented. This way, they can be tested in their ability of simulating and predicting human data. Since these models provide different accounts for explaining the experimental findings, the most important models will be introduced before the experimental data will be discussed. The aim is to emphasize the differences between the models. As the purpose of this chapter is to provide insight in factors that play a role in speech processing, I will not treat excessively how each model accounts for the selected data. Rather, I will try to provide a deeper understanding of how the different assumptions are related to different interpretations of the experimental results. As will be seen, when these different perspectives are taken into account, it is possible to appreciate and recognize the aspects on which there is actually fairly broad consensus. Therefore, I will conclude this chapter with summarizing how the experimental data reflect some general properties of human speech processing. First, however, the general framework within which psycholinguistic research on spoken word recognition takes place will be described.

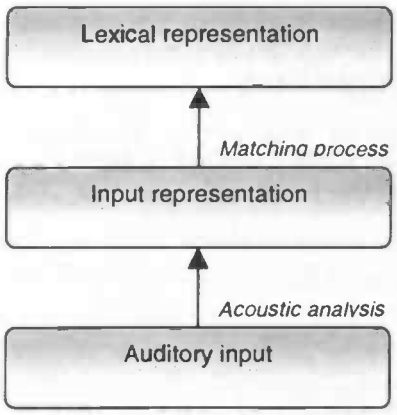


Figure 5.1 Processing stages and representations involved in auditory lexical access.

5.1 Spoken word recognition

In the perception of connected speech, word recognition plays a central role in the processes by which acoustic waveforms are converted into a representation of the meaning of utterances. It is generally assumed that speech recognition involves progressive stages of abstraction, going from a representation of the acoustic signal, to low-level linguistic or perceptual units, to lexically based representations and ultimately to syntactic and semantic properties of the spoken input. A commonly made distinction is between

(1) processes that are involved in deriving an initial, form-based representation of the auditory input (*acoustic analysis*), and
(2) later stages involved in accessing representations of specific lexical candidates (*lexical access*) that match the input representation in order to identify or recognize the lexical items contained in the speech signal (*matching process*).

This is illustrated in figure 5.1.

There is general agreement regarding this abstract processing framework, but there is considerable disagreement regarding for instance the nature of the representations that are involved in each of these stages, and the interaction between different levels of representation during processing. Within this chapter, the following specific issues will be addressed:

- *Representational units*: The nature of the prelexical representation of the speech input that contacts the mental lexicon (section 5.3).
- *Lexical access*: The process by which input representation are mapped onto representations of lexical form, in particular, whether lexical items are compared successively or simultaneously with delayed commitment to a single lexical item (section 5.4).
- *Lexical competition*: The consequences of the nature of this mapping process on the activation of lexical representations in the process of word recognition (section 5.5).
- *The direction of information flow*: The influence that activated lexical items have on prelexical processing (section 5.2.1 and 5.6).

5.2 Models of spoken word recognition

The computational models that exist can be distinguished in how they account for the issues just described and related issues. Some of the most influential models will be introduced in this section.

5.2.1 Interaction "versus" autonomy

A distinction that is often made between models of psycholinguistic processing - or more generally, perceptual processing - is based on whether processing at different levels is viewed as interactive or autonomous (cq. modular). However, dependent on the specific field of interest these terms have different meanings (see also Boland and Cutler, 1996).

In the case of syntactic processing, for example, it relates to whether syntactic choices are made with the benefit of semantic knowledge. One theory states that syntactic alternatives are constructed in parallel with the constraints of lexical specifications, and a single representation is selected by the semantic system, using principles of referential support, a priori plausibility, etc. (Altmann and Steedman, 1988). Note that a distinction is made between *generation* and *selection*: there is a bottom-up generation of alternatives, with selection of a single structure left for a later stage of processing. This model is classified as (weakly) *interactive*, and is contrasted with strongly interactive models which generate only the most plausible structure(s), and with Frazier's (1987) autonomous model, which generates only the simplest structure. In both instantiations of interactive models, syntactic candidates are evaluated in parallel (i.e., multiple outputs are produced), whereas in Frazier's autonomous model only a single syntactic structure is considered at a time.

Separating generation from selection processes can also be seen in models of word recognition. For instance, the "checking model" (Norris, 1986) is a model for *visual* word recognition. Here, a candidate set, compiled on the basis of partially analyzed perceptual information, is generated. This set is continually updated as the perceptual analysis is refined. In the meantime, selection processes begin with checking the candidates for compatibility with the sentential or semantic context constructed so far in the recognition process.

Because of the temporal nature of speech, an initial rough analysis of the entire word is generally seen as inappropriate for models of *spoken* word recognition. While written language can be perceived in parallel for as long as is required for processing (at least at the level of individual words), spoken language is sequentially ordered: word beginnings arrive temporally prior to word endings, word length can not be initially apparent, and word boundaries are usually not obviously present in connected speech. In addition, the speech input is transient. Only a small amount of speech can be retained in the auditory system in an unanalyzed or echoic form. Therefore, it seems that the processing infrastructure for speech perception must operate rapidly if connected speech is to be processed efficiently.

The interaction versus autonomy dichotomy serves to contrast different aspects of processing. In the parsing literature, use of higher-level information to resolve lower-level decisions constitutes interaction. In word recognition, a process is not taken to be interactive unless higher-level information actually *affects* the way that alternatives are generated within the system. This way, certain candidates could be ruled out irrespective of their compatibility with bottom-up information. Autonomy implies that processing operations at a given level proceed in the same way irrespective of what might be deducible from higher-level considerations. This kind of autonomy is comparable with Fodor's (1983) arguments for modularity in mental processing: "a system is autonomous by being encapsulated, by not having access to facts other systems know about". In autonomous models, a distinction can therefore be made between generation and selection of word candidates.

Therefore, when placing models of speech recognition along the interaction versus autonomy dichotomy, this pertains to the nature of information flow between different levels of processing. Within interactive models, higher levels of processing can *alter* processing on lower levels through feedback connections. In autonomous models, processing at each level operates independently of other (higher) levels and each level delivers its output independently. Decisions are based on reading out one of the "outlets". Or, the output of a level gets integrated with outputs of other levels to make decisions or to form one coherent percept. In fact, most autonomous models integrate all available - independently estimated - sources of information to reach one optimally coherent interpretation of the perceptual scene.

Psycholinguistic models of spoken word recognition can be characterized in how they are placed along the interaction versus autonomy dichotomy. The discussion of the evidence in favor and against this aspect of processing will take place throughout the chapter, though mainly in section 5.6.

5.2.2 Trace model

An example of a highly interactive model is the *Trace* model (McClelland and Elman, 1986) which is based on an interactive activation and competition model of processing (IAC), and is inspired by a model of visual word recognition (McClelland and Rumelhart, 1981). In Trace, information processing is carried out in a network consisting of a large number of interconnected units where different classes of units represent acoustic-phonetic features, phonemes or words. Each of these classes of units constitute one level of processing. The units within one level are connected through bidirectional inhibitory connections (*competition*) and between levels there are bidirectional excitatory connections (*interaction*).

This process is illustrated in figure 5.2. Each unit stand for a *hypothesis* about the input being processed, and is working *continuously*[52] to update its own activation on the basis of activations of other units - whose activation level have exceeded some threshold activation level - to which it is connected. The IAC process allows each hypothesis both to constrain and be constrained by other mutually consistent or inconsistent hypotheses. The model is called the Trace model because the network of units forms a *dynamic processing structure* called *"the Trace"*, which *serves at once as the perceptual processing mechanism* and *as the system's working memory*.

Trace is able to simulate a huge amount of empirical findings within one coherent framework and is representative for the class of interactive models. It is important to note here that the feedback connections result in an influence of higher order units on lower level units thereby *altering* the information processing at these lower levels. Thus, *no distinction is made between generation and selection processes.*

Furthermore, to account for experimental results on either word recognition, phonemic decision-making, and feature detection or discrimination, it is assumed that the *information* that is represented by the different units *at each level of processing is explicitly accessible* for the purpose of making decisions.

Regarding lexical segmentation, Trace uses its knowledge of the lexicon to parse sequences of phonemes into words, and to establish where one word ends and the next one begins when cues to word boundaries are lacking. Segmentation involves a competitive process among potential words and the actual word defined sequentially by the input. Lexical hypotheses wax and wane depending upon their match with the input and competition from other lexical hypotheses.

[52] Of course, during simulation with a computational implementation of the model, the underlying continuity is approximated by taking discrete time steps, where at each time slice all activation levels are up-dated.

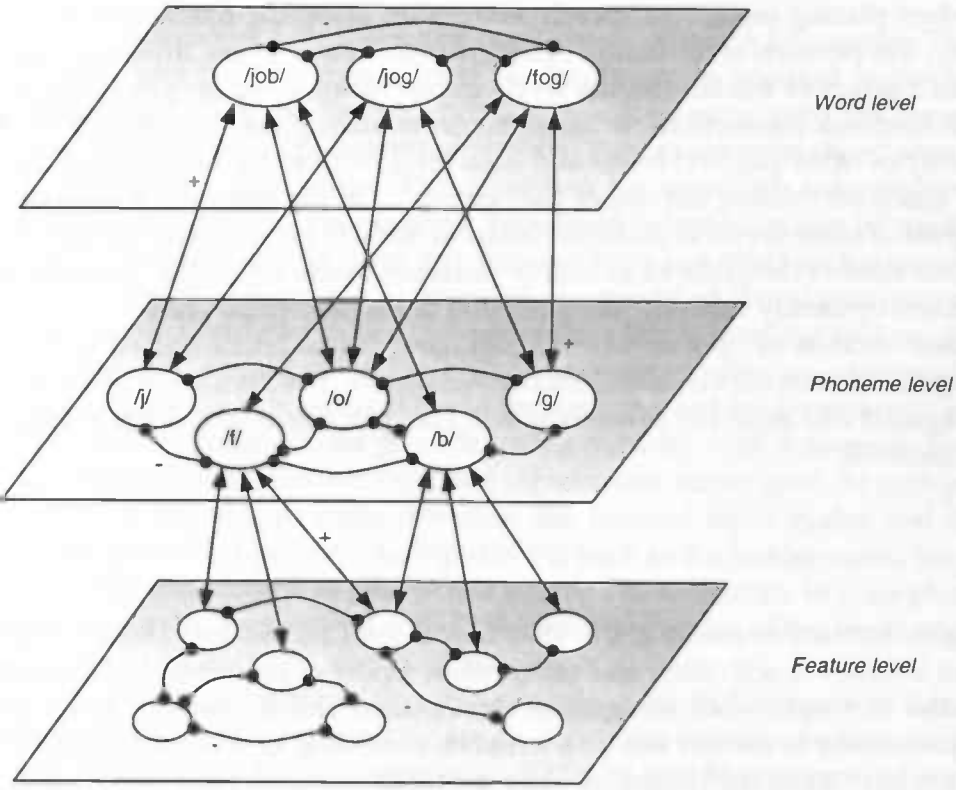


Figure 5.2 Schematic representation of the interactive *Trace* model (McClelland and Elman, 1986). This model is implemented as an interactive activation and competition network. In spoken word recognition, three levels of processing can be distinguished where processing units that (roughly) correspond to features, phonemes and words compete with one another within one level. Between levels bidirectional excitatory connections exist between the processing units. The amount of excitation or inhibition between units is determined by the degree of activation.

5.2.3 Shortlist model

The connectionist model *Shortlist* (Norris, 1994b) is an example of an autonomous model: processing at one level is unaffected by processing at another. Information-processing (i.e., lexical processing) at subsequent levels is entirely bottom-up. Furthermore, different levels (can) have different processing operations. In Shortlist, a distinction is made between a generation stage and a selection stage. In spoken word recognition, multiple lexical candidates that are roughly consistent with the bottom-up input, and for which the bottom-up activation is sufficiently high, are activated during the generation stage, thereby forming a short list of hypotheses that serve as input to the selection stage. The *degree of activation* of the different candidates depends on the *similarity* - according to a phonemic-featural description - *with stored lexical representations*. This activation level builds up over time.

In contrast to Trace, there is no lateral inhibition at the phonemic-featural level. According to Norris, *the absence of inhibitory connections at the prelexical level is essential in a bottom-up system*, as inhibition at this level would have the effect of producing categorical decisions that are difficult to overturn for other levels. As a result, information vital for the optimal selection of a lexical candidate could be lost. During perceptual processing, the ambiguous nature of the input should be preserved, i.e., no decisions ought to be made. Only when the circumstances require such early decision-making (as, for example, in the pho-nemic-decision tasks that will be discussed in later sections), there is a competition, through inhibition, between phonemic descriptions, but this occurs at a decision-level (see for example the Merge model (Norris et al., 2000) where a similar distinction between perceptual pro-cessing and decision making is made).

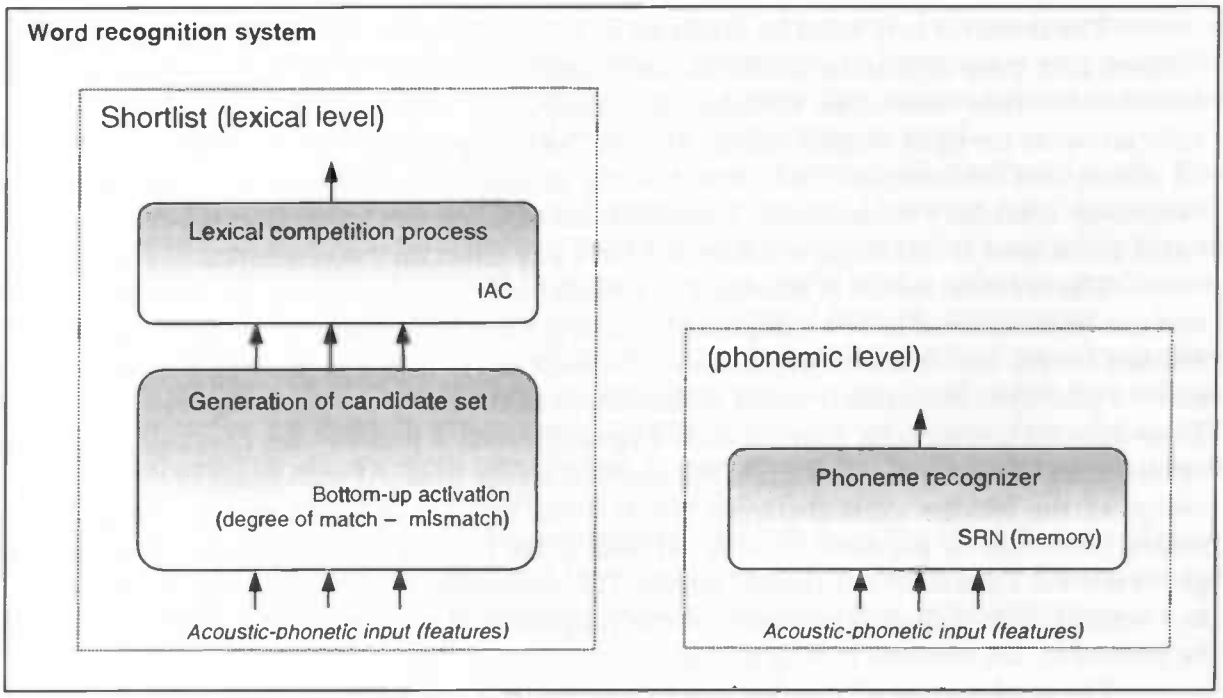


Figure 5.3 Schematic representation of the autonomous *Shortlist* model of spoken word recognition (Norris, 1994b). Regarding lexical processing, a distinction is made between a generation stage and a selection stage. An initial candidate set is generated on the basis of bottom-up input - consisting of acoustic-phonetic feature bundles - alone. Selection among this multiple lexical candidates takes place through a process of competition within the generated candidate set, but this process has no influence on generation. The list of candidates is continuously updated in the light of new information. The recognition of phonemes is not influenced by lexical processing, and is also not required for lexical processing. In order to make phonemic decisions, two sources of information are available: one based on prelexical and one based on lexical processing.

During spoken word recognition, it is only at the lexical level that decisions ought to be made.[53] When lexical candidates share segmental information, they enter into competition with one another at the selection stage via lateral inhibition within an interactive activation and competition network.[54] The degree of inhibition between multiple candidates depends on the degree of overlap. Like in Trace, the competition mechanism also results in lexical segmentation. However, in Shortlist, this lexical "knowledge" (related to the "*possible word constraint*") is combined with an explicit incorporation of the so-called *Metrical Segmentation Strategy* (MSS) (Norris et al., 1995). The MSS assumes that listeners (in English[55]) hypothesize word boundaries for syllables containing strong vowels. So, a distinction is made between strong and weak vowels: a word containing a strong, initial vowel that is incorporated into the short list is favored during the lexical competition stage to permit modeling of lexical segmentation and competitor effects.

[53] In this context, *decision* refers to nonconscious processes, not a conscious decision on the part of the observer. Also, the term decision-level should not be interpreted as a separate level of processing within the perceptual processing system dealing exclusively with making decisions about perceptual input. Rather, it refers more to attentional processes of selection leading to contrast enhancing or inhibitory processes, which are related to specific task demands. Such a decision-level is therefore *constructed on the fly*, and exerts its influence after an initial analysis of the perceptual input has taken place.

[54] It should be noted that the units in an IAC network (as in other localist connectionist networks) represent alternative hypotheses about the current input, and the connection weights between the units encode the constraints that govern what combination of hypotheses constitute good solutions or interpretations. This process can therefore also be implemented by any other connectionist constraint satisfaction procedures, such as a Hopfield net or a Boltzman machine.

[55] Norris and colleagues stress that listening is language-specific. Therefore, the exact MSS differs between languages, but it seems that there are characteristic patterns of intonation (stress patterns and/or rhythmic patterns) for many languages that provide a valuable means for additional segmentation strategies.

The model is very *sensitive to bottom-up mismatch information*. The list of potential candidates, and their activation levels, is continuously updated in the light of new available information. This continuous updating in combination with the competition process naturally accounts for right-context effects, i.e., the "re-interpretation" of information in the light of subsequent information. The correction for mismatch information is accomplished via bottom-up inhibitory connections. Top-down information can be used to select among activated candidates at this stage, but does not have any effect on the generation of candidates.

The Shortlist model is intended as a model of lexical processing in word recognition, and can be considered to be a component in a larger modular system of word recognition. To account for the results found in phonemic-decision tasks, an additional autonomous component (a phoneme recognizer) could be implemented by for example a recurrent network.[56] Time-delayed connections provide such a network with a memory for previous processing operations and enable it to integrate information across time. At each point in time, the activation of the hidden units (between the featural input nodes and the phonemic output nodes) generated by previous input is fed back to the hidden units. Such networks therefore give a natural account of left-context effects. The sensitivity to prior phonemic context results in a sensitivity to statistical and phonotactic properties of the input and such properties will be learned by the network during training.

The combination of Shortlist and this additional component is consistent with models that are specific for explaining data on phonemic-decision making, such as *Race* (Cutler and Norris, 1979) and *Merge* (Norris et al., 2000). In fact, Shortlist can be seen as a computational implementation of lexical processing in these latter models. The distinction between Race and Merge is that in Race, phonemic decisions are either based on a prelexical or a lexical representation, whereas in Merge, phonemic decisions are made by integrating knowledge obtained from both representations.

5.2.4 Cohort model

A comparable model is the *Cohort* model of Marslen-Wilson and Welsh (1978). One of the basic assumptions of this model is that the first sound(s) of a word is used to determine which words will be an initial candidate set. This reflects a reliance on word-initial information and pertains to the idea that *word beginnings are vital to fluent recognition*. The initial cohort is based on the basis of acoustical input alone (bottom-up).

In an earlier version of the model, it was assumed that as each successive phoneme arrives, words are eliminated from the cohort immediately if the new phoneme fails to match the next phoneme in the activated word (using phoneme-to-word inhibition), or on the basis of semantic constraints. In a later version, a less strict mapping is required: candidates are *activated to the degree that they match the acoustic-phonetic input*. This way, initially ambiguous word onsets can be overcome. Furthermore, there is a *continuous mapping of featural information onto the lexical level*. Word recognition occurs immediately as soon as the cohort has been

[56] Notice that the incorporation of an additional autonomous component explicitly dealing with the recognition of phonemes reflects that this is not considered to be part of the process of spoken word recognition. Within many experimental tasks subjects are asked to make phonemic decisions. Though the purpose of these tasks is to tap into normal speech recognition processes, Norris and colleagues consider this as only partly true. The demands that are placed on subjects by attending to phonemic units and making decisions about them is different than those of normal speech recognition. In a bottom-up model, *no decisions* are made about phonemes, and within Shortlist there does *not* exist an *intervening phonemic layer*. Furthermore, it is assumed that explicit knowledge of the sounds of a language can be used to make phonemic decisions. Prelexical processing therefore means in this context: processing in which lexical knowledge is not used, since attention is focused on the acoustic-phonetic input, and not: processing that takes place before lexical access in normal speech recognition (see also section 5.3.2).

reduced to a single member (i.e., after the *isolation point*[57] of a word). This has been known as the basic *immediacy assumption*. In this sense, *word segmentation is a byproduct of recognition*: the word boundary is implicit at the offset of the speech stream that defines a word in that the onset of the following word is a default of recognizing the previous word.

Activation of any lexical hypothesis is evaluated with reference to other activated words. This pertains to the competitive nature of lexical access and selection. However, there are *no direct effects of competition* among simultaneously activated words. Evidence for a particular lexical pattern accumulates during temporal processing of the signal and a selection mechanism must determine when one candidate has sufficiently diverged from its competitors.

Furthermore, there is an active role of context in reducing the cohort to a single competitor. *Context biases selection by positively activating patterns that are already active from bottom-up input*. In earlier versions, it was a viable means to exclude candidates from the cohort (through inhibition of inconsistent candidates). It now serves to "pull" contextually appropriate words further from their competitors in the cohort.

Finally, it is assumed that word recognition can influence the identification of phonemes in a word only after the word has been recognized. The Cohort model therefore explains the results on phonemic-decision tasks as based on postlexical representations.

5.2.5 Logogen model

An integrative, autonomous model is the *logogen* model (Morton, 1970), which is a *descriptive* model of word recognition. In this model, words are represented in memory as counters, or logogens. Incoming stimulus information is accumulated in the logogens, and identification of the stimulus occurs when the amount of evidence in a word's logogen exceeds a certain threshold. Thus, different sources of information are being processed independently of one another and then integrated to form one percept. Information-processing from auditory input to a word's representation in memory[58] is in one direction. Higher order processes only have their influence at the integration stage, where they just add evidence from another source of information. This model distinguishes from the previous models in that only one lexical entry will surmount a recognition threshold and be effectively accessed.

Within the logogen model, high-frequency words differ from low-frequency words in that their resting activation level is higher, i.e., their thresholds for recognition are lower. It can therefore account for *word frequency effects* (e.g., faster recognition for high- compared to low-frequency words). It can also account for *repetition priming effects*. When a word is primed through previous presentation, this leads to some amount of residue activation for a short duration of time (or, in other words, a temporary lowering of thresholds) which results in faster access and therefore faster recognition of this word with repeated presentation.

[57] The isolation point of a word corresponds to its *uniqueness point*, i.e., the point at which the word can be uniquely identified.

[58] The representation of memory should not be interpreted as information being stored and available for use in subsequent tasks. Rather, memory is that a word comes to attract more than its fair share of the perceptual evidence entering the system from a stimulus.

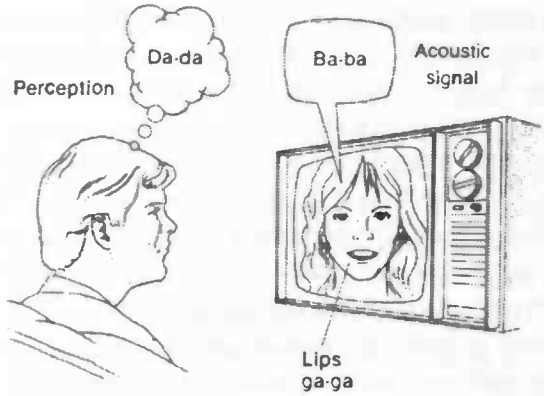


Figure 5.4 Audiovisual speech perception. The woman is making the lip movements for /ga-ga/, but the actual sound is /ba-ba/. The listener, however, reports hearing the sound /da-da/. (From Goldstein, 1990).

5.2.6 Fuzzy Logical Model of Perception

Within the *Fuzzy Logical Model of Perception* (FLMP) (Massaro and Oden, 1980, 1995; Massaro and Cohen, 1991) strict independence in the basic perceptual processes is being maintained, while allowing contextual information to have its influence on the final outcome of perceptual processing through integration of information. A clear example showing that different sources of information both exert their influence on the perceptual outcome, comes from cross-modal studies where visually presented information has an influence on the perception of simultaneously presented auditory stimuli (see figure 5.4). For example, visual presentation of a person saying /ga/ combined with the auditory stimulus for /ba/, leads to the auditory percept of /da/, an effect that is known as the *McGurk effect* (McGurk, 1968).

The FLMP is a *feature-integration model*. It assumes that sensory systems transduce a physical event and as a result make various sources of information (*features*) available. Perceptual effects are processed in accordance with a general algorithm where the following three operations are carried out: (1) feature evaluation, (2) feature integration, and (3) decision. A schematic representation of this process is depicted in figure 5.5. The (continuously valued) features are evaluated and matched against prototype descriptions in memory by a process that integrates individual feature values *according to the specifications of the prototype*. The evaluation of an information source provides information about the strength of alternative interpretations, rather than just all-or-none categorical information, as claimed by "categorical perception" theory (see, for example, section 5.3.2.2).

After feature integration, an identification decision is made on the basis of the *relative goodness-of-match* of the stimulus information with the relevant prototype description (the probability of a response is the merit of that alternative, relative to the sum of the merits of all relevant alternatives). This leads to a prediction of the proportion of times the stimulus is identified as an instance of a certain prototype, or a rating judgment indicating the degree to which stimulus and category match. Thus, the FLMP is *deterministic at feature evaluation and integration processes*, and becomes *stochastic only at the decision process*. It can account for *feature trade-offs* and provides a very *good quantitative fit to categorical perception* results (without assuming categorical perception!).

A prediction of the FLMP is that *the impact of one source of information on performance becomes larger with increases in ambiguity of other available sources of information*. As information sources are evaluated independently, the integration process insures that the least ambiguous sources have the most influences on the judgment. It gives a natural account of the integration of bottom-up and top-down sources of information, which are processed over time.

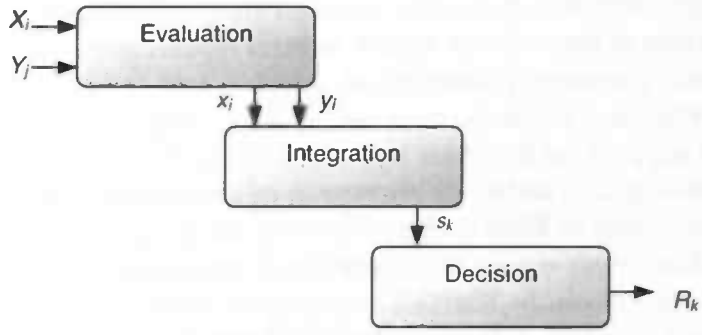


Figure 5.5 Schematic representation of the *Fuzzy Logical Model of Perception* (Massaro and Oden, 1980), consisting of three stages of processing that are involved in perceptual recognition: feature evaluation, feature integration and decision-making. These stages are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. The sources of information are represented by uppercase letters (X_i and Y_j). The evaluation process transforms these sources of information into psychological values (x_i and y_j). These sources are then integrated to give an overall degree of support for a given alternative s_k. The decision operation maps this value into some response, R_k, such as a discrete decision or a rating.

Therefore, FLMP *can account for context dependency while maintaining strict independence in the basic perceptual processes* (as is the case in the *Ganong effect* (Ganong, 1980), see also section 5.6.1).

However, the FLMP model is a mathematical model, where a parameter estimation procedure is used (for example by means of a root-mean-square deviation (RMSD) measure) to "explain" obtained results. Though the FLMP provides a quantitative fit to many experimental results (even for the performance of individual subjects), it *does not describe the dynamics of processing* that leads to the obtained results.

Though it is claimed that the independence that is assumed in the summation of different sources of evidence at the feature integration stage also reflects architectural independence, it is actually neutral in this respect. As McClelland (1991) points out, FLMP only predicts, from looking at behavioral data, that the system *behaves as if* different sources of information are processed independently. It therefore describes the system at another level of description than models such as Trace, Shortlist and Cohort.

5.3 Representational units

The process of spoken word recognition requires the listener to relate an acoustic-phonetic analysis of the speech signal to stored representations of lexical form in the mental recognition lexicon. A basic issue concerns the unit of processing that is extracted from the speech signal to activate a lexical representation in the process of lexical access.

Linguists typically describe words as a sequence of phonemes, which in turn can be distinguished by a set of distinctive acoustic-phonetic features. Phonemes are discrete units. It is clear that during the process of continuous speech recognition, there are no discrete segments present in the physical signal corresponding to individual phonemes. In addition, the *features* that signal the presence of a particular phoneme vary within different (syllable or word) contexts. Linguists argue that such variations are not random. In fact, within a given language, there are phonological rules that can predict such allophonic variations by taking the preceding and or following context into account (see section 2.2). The variations reflect predictable coarticulation effects that are the direct consequence of (the limitations of) articulatory processes. For example, when a vowel is followed by a nasal, this nasalization is already present in the realization of the vowel. These variations are therefore reflected in the *acoustic realization* of phonemes, so they differ among contexts. Furthermore, they differ within and between speakers.

Within the field of psycholinguistics, it is generally believed that understanding spoken language requires a transformation of the acoustic signal. Several representational levels have been proposed, such as acoustic, phonetic, phonological, syllabic, and lexical levels. A distinction between acoustic properties and phonetic features has been supported by many psychological data, but there is also support for the other hypothesized codes.

A particular issue is the psychological validity of phonemes as a representational unit intervening between the lexicon. According to Klatt (1980) phonemes are not psychologically realized during speech perception. Klatt even argues that there are no intermediate representations - or abstractions of the signals - at all. In Klatt's *Lexical Access From Spectra* (LAFS) model, spectral templates serve to activate lexical representations through a lattice of possible templates.

Warren (1976) also concludes that "...while phonemes are constructs useful for transcribing and analyzing, they are without direct perceptual basis... phonemes seem to have no direct relevance to perceptual processes leading to the comprehension of speech". Marslen-Wilson and Warren (1994) argue instead for a direct access feature-based mapping. However, there are also models, such as the Trace model (McClelland and Elman, 1986) in which a phonemic layer is explicitly incorporated as an intervening layer. Lower level cues, such as feature information, are integrated to identify higher order units that constitute the input to the word units. Others argue for units larger than phonemes that represent structure at a level that spans multiple phonemes or feature clusters (e.g., syllables (Greenberg, 1996)).

5.3.1 Acoustic versus phonetic levels of analysis

The distinction between acoustic and phonetic levels of analysis has been supported by many experimental data.

5.3.1.1 Dichotic fusions

For example, Cutting (1976) conducted a parametric investigation where acoustic properties such as relative onset synchrony, intensity, and fundamental frequency were manipulated. It was tested how these manipulation affected different kinds of *dichotic fusions*. These different fusion types can be illustrated by describing how each could be involved in the perception of /da/:

(1) Presenting a single stimulus to each ear, leads to a single percept /da/. Localization of the stimulus depends on various factors (e.g., relative intensity and timing of the /da/ in each ear);
(2) *Psychoacoustic fusion*: presenting /ba/ to one ear simultaneously with /ga/ in the other often leads to the percept /da/ (70 %). This can be explained by an averaging of energy in the F2 region, i.e., /ba/'s sharply rising formant with /ga/'s sharply falling produces /da/'s relatively flat F2;
(3) *Spectral fusion*: presenting F1 of /da/ to one ear, and F2 to the other also induces the /da/ percept;
(4) *Spectral/temporal fusion* or *duplex perception*: presenting F1 and F2 dichotically, with different acoustic properties, leads to one phonemic percept /da/, seemingly produced by two sources (see also section 4.10).

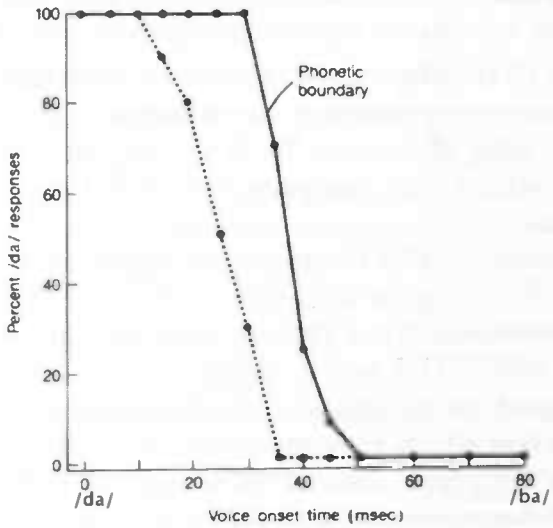


Figure 5.6 Adaptation effects. Solid line: the percent of sounds identified as /da/ as voice onset time is varied from 0 to 80 ms. Dotted line: the percent of sounds identified as /da/ after adaptation to /ba/ for two minutes. The phonetic boundary's shift to the left indicates that /da/ is heard less often after adaptation. (From Eimas & Corbit, 1973).

(5) *Phonetic feature fusion*: /ba/ paired with /ta/ results in a percept of /da/ or /pa/. Here, phonetic features as voicing and place appear to get recombined. It therefore reflects a more abstract level for fusion, as the derived percept could not be constructed by any averaging of formants of /ba/ and /ta/;

(6) *Phonological fusion*: /da/ paired with /ra/ leads to the percept /dra/, consistent with phonological rules, or phonotactic constraints.

To account for how the fusions behaved as a function of testing conditions, Cutting argued that at least three different levels of analysis were needed. For example, regarding the sensitivity to amplitude imbalance, it was found that presenting the dichotic components at different intensities inhibits fusions of the last two types, but has no effect at all on spectral or spectral/temporal fusion. The following levels of analysis were proposed:

(1) an *acoustic* level that is responsible for sound localization
(2) an *acoustic* level where psychoacoustic, spectral and spectral/temporal fusion are grouped
(3) a *linguistic* level including phonetic feature and phonological fusion types

5.3.1.2 *Selective adaptation effects*

Eimas and Corbit (1973) investigated different levels of analysis of speech by studying *selective adaptation* effects. Adaptation is a robust phenomenon in the speech literature. Within the adaptation paradigm, an adaptation stimulus - typically an endpoint stimulus from a certain continuum, for example, from /da/ to /ba/ - is repeatedly presented resulting in a boundary shift of two phonemes along the speech continuum. This means that when - in the test condition - listeners have to identify stimuli forming a continuum, the adaptor produces fewer labeling responses of the adaptor's category thereby shifting the category boundary away from the adaptor (see figure 5.6).

Eimas and Corbit showed that adaptation effects could also be obtained with adaptors that were *not* members of the test continuum, but shared important properties with the test items. For example, repeated presentation of /ba/ - sharing *voicing* features with /da/ - reduced the report of /da/ on a /da/-/ta/ continuum in a phonemic identification task. This reflects an adaptation effect at a *phonetic feature* level of analysis. In an investigation conducted by Tartter and Eimas (1975) adaptation effects with (nonphonetic) *single* formants were found, reflecting an *acoustic* level of analysis.

Also using the adaptation paradigm, Sawusch (1977) illustrated the abstractness of higher levels of analysis. In one condition, adaptors having the same pattern of formant transitions as the endpoints of the continuum were synthesized, but all three formants were displaced along the frequency scale by 1.5 critical bands. This was to ensure that the shifted adaptors could not produce effects at a level based on information directly available in a neural spectrogram. Nonetheless, reliable adaptation effects were produced, about one half the size of the effects caused by the continuum endpoints themselves. Hence, the shifted adaptors appear to have had their effect at a level that represents the *integration of information from multiple formants*, but that is *not tied to particular frequency regions*. Monaural presentation of these shifted adaptors in the ear *contralateral* to the monaurally presented test items produced the *same size of adaptation effects* as with the ipsilateral adaptation test. Sawusch therefore concluded that the shifted adaptors were operating on a binaurally driven, and therefore more central, representation.[59] In contrast, the continuum endpoints themselves, when used as adaptors, produced smaller effects contralaterally than ipsilaterally, indicating that part of their effect was based on monaurally driven representations.

In a more recent investigation, Samuel and Kat (1996) studied the nature of such adaptation effects in more detail. They did not only analyze the identification data, but also analyzed the effects on RTs. Within the adaptation literature, two effects are typically found: RT shifts and criterion shifts. A possible interpretation is that RT shifts occur when lower level units are *fatigued*, such that repeated exposure leads to a reduced (i.e., lower and slower) responsiveness of representations. On the other hand, a boundary shift without RT shifts is due to a higher level *criterion shift*. Of course, an alternative interpretation could be in terms of criterion shift models that can *also* produce RT shifts. Samuel and Kat reasoned that if one testing situation reliably produces RT shifts, and another situation reliably does not, it can safely be inferred that different mechanisms must be attributed to the two cases. To test this, they used different types of adaptors:

- a baseline condition: /a/, sharing no critical features that can distinguish between /b/ and /d/ in the test continuum
- the standard continuum endpoint adaptors: /ba/ and /da/
- acoustic derivatives of each endpoint: only F2 and F3 (F2F3:B and F2F3:D) therefore containing most of the critical acoustic information differentiating /ba/ from /da/ but carrying very weak phonetic information[60]
- phonetic analogs: the voiceless counterparts /pa/ and /ta/
- alternating sequences of each of the last three types: /ba/-/da/-/ba/-/da/..., F2F3:B-F2F3:D-F2F3:B-F2F3:D..., /pa/-/ta/-/pa/-/ta/...

[59] Remember from chapter 3, that binaural information in the *central* auditory system is already available in the Superior Olivary Colliculus, far before it reaches the auditory cortex.
[60] The combination of F2 and F3 is insufficient to form a phonetic percept. It is necessarily for F1 to be present.

Experimental condition		
Adaptation	**Test**	**Results**
Binaural: - F2F3 - /pa/-/ta/	Binaural	- Reliable identification shifts on /ba/-/da/ continuum, slower RTs - Same size of identification shifts, no RT effect
Conclusion: The distinction between acoustic and phonetic adaptors is supported by the dissociation in RT data.		
Monaural: - F2F3 - /pa/-/ta/	- Ipsilateral - Contraleral - Ipsilateral - Contraleral	- Reliable identification shifts, accompanied with RT effect - No difference: same size of adaptation effects - Reliable identification shifts, *not* accompanied with RT effect - Same size of adaptation effects
Conclusion: Evidence for two binaurally driven levels of representation: one subject to RT changes and one not.		
Monaural: - F2 or F3 - F2F3	Binaural	- Single-formant adaptors can produce small, but reliable, boundary shifts - Adaptation effect as reflected in identification shifts significantly larger than sum of individual formants adaptation; also difference in RT effect, but not significant
Conclusion: Identification shifts and RT changes indicate that the level affected by the complex acoustic F2F3 adaptor is indeed integrative, i.e., integrating information from the individual formants.		
Dichotic: F2 / F3 Binaural: F2F3	Binaural	- The effects produced by the dichotically presented F2/F3 adaptor are identical to those for the binaural F2F3 adaptor
Conclusion: The different presentations of the adaptors affect the same integrative level of analysis, thereby supporting that the two-formant acoustic stimuli integrate information from the two ears and from the two formants.		

Table 5.1 Summary of the results obtained within the selective adaptation experiments performed by Samuel (1996).

Furthermore, they used binaural, monaural and dichotic adaptation procedures, and ipsilateral and contralateral testing conditions. The obtained results again indicated that at least three different levels of analysis were involved: (1) a "simple" *acoustic* level, (2) an *integrative acoustic* level, and (3) a more *abstract categorical* (possibly phonetic) level. For example, the F2F3 combination presented to one ear led to the same adaptation effects when presented either ipsilateral or contralateral to the adaptor ear in the test condition. This could not be explained by the adaptation effects of F2 and F3 when presented in isolation. It therefore reflects an *integrative acoustic, binaurally driven* level of representation.

This was confirmed by the fact that dichotic presentation of the F2 and the F3 adaptors led to the same adaptation characteristics when testing the F2F3 combination compared to intact, binaurally presented F2F3 adaptors. The most important results of the Samuel (1996) investigation are summarized in table 5.1. The derived properties of the different levels in terms of their locus, the nature of the adaptation shifts, and their stimulus domains are listed in table 5.2.

Regarding the third level, it is also suggested by prior work that it is indeed processing *phonetic* information. Sawusch & Jusczyk (1981) found, for example, that in a contrast paradigm, that should rely on criterion shift effects, *the perceived phonetic quality of a stimulus dominated the results.*[61]

However, there are also results suggesting that this level may *not* be speech specific (Diehl, 1976). Diehl showed that a nonspeech sawtooth wave that sounded like a plucked violin string produced reliable adaptation effects on a /pa/-/wa/ continuum. This effect was undiminished under contralateral listening conditions.

[61] It is important to realize that the synthetic stimuli that are used form an abstraction of normal speech, consisting of the trajectories of formant bands, defined by a formant peak and a certain bandwidth under which harmonic components are present. It seems that these stimuli can only produce reliable adaptation effects at higher levels of abstraction if they are truly interpreted as speech.

Properties	Level I	Level II	Level III
Central vs peripheral locus	Peripheral (monaurally driven)	Central (binaurally driven)	Central (binaurally driven)
Nature of adaptation shift	Fatigue (?)	Fatigue or criterion shift (RT)	Criterion shift (non-RT)
Stimulus domain	"Simple" acoustic (neural spectrogram)	Integrative acoustic	Categorical (phonetic) ?

Table 5.2 Overview of analytic levels of processing

Furthermore, the effect was *just as strong* when the test series was created in a *whispered* form, greatly reducing the acoustic match between the (periodic) adaptor and the (aperiodic) endpoint, indicating adaptation effects at a highly abstract level, but not necessarily speech-specific.

A possible explanation for these results is that the third level of processing reflects the responsiveness of (groups of) neurons that are sensitive to a specific FM *pattern*. In phonemes, such abstract patterns correspond to formant transitions that allow for the distinction between consonants. In the case of nonspeech stimuli, a similar pattern of FM may therefore lead to the activation of the same groups of neurons. This might also explain why the adaptation effects for whispered speech and normal speech are equally strong, as *the detection of formants does not depend on - though it is facilitated by - the presence of voiced components*. This is consistent with many of the findings described in chapter 3 and 4, and again indicates the abstract representation of relevant speech features.

Therefore, the results all seem to converge on a simple acoustic, an integrative acoustic and a more abstract (possibly phonetic) level. Interestingly, the results of Cutting (1976) map very well on the results obtained by Samuel and Kat (1996), even though the results were obtained by taking different approaches. In addition, the results are consistent with the results reported in previous chapters.

Of course, a fully specified model will include a description of what each level does, including the sort of properties as listed in the above table. It should also clarify the time course of effects at each of the levels and the arrangement (serial, cascade, parallel) and communication between levels (uni- or bidirectional). The long-term goal of this kind of research is to delineate all of the transformation that complex sounds undergo, beginning with a "neural spectrogram" and culminating with the codes used to achieve lexical access.

5.3.2 Phonemes as an intervening level of analysis

Based on the sensitivity to abstract (phonetic) features as reported in the previous section, it can be concluded that some kind of processing takes place on phonetically related representations. However, this still does not imply that, during the process of normal speech recognition, such phonetic features are integrated into higher order phonemic "units" before lexical access occurs. Within phonetic decisions tasks this issue is indirectly addressed.

5.3.2.1 Phonetic versus phonological codes

In the phoneme monitoring (PM) task listeners have to detect a previously specified *target* phoneme in spoken input and have to respond by pushing a button if and when the target occurs, i.e., respond as soon as they hear a sound in a (list of) word(s), or when they hear a (word-initial) prespecified target in a sentence. For example, listeners have to respond as

soon as they hear the target phoneme /b/ as in *brain*. Reaction times (RTs) are measured (for a review, see Connine & Titone, 1996).

Variables that can be manipulated are for example: lexical status (words vs. nonwords), word frequency (low vs. high frequency words), the number of targets to be monitored, the presence of distracting (phonetically similar) phonemes, the predictability of a word in a sentence context, or the influence of attention (e.g., by requiring subjects to perform a concurrent task).

Within the PM paradigm, one of the questions is whether listeners PM responses is based on a *prelexical* (*phonetic*) or a *postlexical* (*phonological*) code. Whereas phonetic codes are computed directly from an analysis of input acoustic information, phonological codes are derived from information available *subsequent* to the activation of higher order (syllabic or word) units. Hence, it is assumed that listeners can carry out the PM task in two alternative ways:

(1) they respond directly to the phonetic code derived from the acoustic information; the lexical item has not yet been identified (though the acoustical signal may already be transformed into a code having linguistic significance), or
(2) they respond to a target phoneme only subsequent to accessing an abstract of the word, that carries the target, which becomes available as soon as a word has been accessed in the mental lexicon; so, they cannot gain access directly to phonetic segments, but decide on the basis of an examination of the phonological representation of the word as a direct result of lexical retrieval.

The reasoning behind the PM paradigm is that if, during the process of spoken word recognition no phonetic segments are being computed, listeners can only respond subsequent to lexical access. If they do compute phonetic segments prior to lexical access, it may be possible (though not necessarily) to respond on the basis of a prelexical, phonetic representation.

A related issue is the contribution of the lexicon in the process of speech recognition. A general finding within many experimental paradigms is that lexical status influences detection times, e.g. listeners can respond faster to words than to nonwords. Also, the time to retrieve a word from the mental lexicon is inversely related to the word's frequency of occurrence in the language, which means that high-frequency words are more easily detected than low-frequency words as a result of faster lexical access. Therefore, the manipulation of variables such as lexical status and frequency should be reflected in PM times, but *only* if listeners respond *after* lexical access.

It has been found that listeners can respond prelexically as soon as they hear the target. In this case, when lexical status or frequency effects of the target-bearing words are manipulated, lexical effects do not occur. However, lexical effects are found when the likelihood of responding on the basis of a postlexical code is increased. This can be achieved for instance by increasing the predictability of the word containing the target in the sentence by providing a semantically biasing context (i.e., the transitional probability is high), which leads to faster lexical access.

Another variable that increases the likelihood of postlexical PM decisions is an additional memory load, which can be achieved by increasing the number of targets to be monitored, or by requiring subjects to perform a secondary task. Monitoring for more than one phoneme (for example, the /b/ and the /p/) and having to choose between them, puts an additional load on subjects in PM tasks. This leads to increasing PM times, and therefore an increased likelihood of lexical access. Having to perform a secondary task also increases the task demands, and therefore the presence of lexical effects. This is also dependent on the nature of the secondary tasks and the resulting time constraints, indicating that attention effects have a greater explanatory validity (e.g., Eimas et al., 1990; Eimas and Nijgaard, 1982; Foss and Blank, 1980).

Additionally, PM times are influenced by the phonetic similarity of the multiple target phonemes, or the target phoneme and the "critical" phoneme. As the number of shared distinctive features increases, so does the RT, and so does the likelihood of finding lexical effects. For example, in an investigation conducted by Foss and Blank (1980), the task was to detect a *word-initial* target phoneme. The critical phoneme is the initial phoneme of the word immediately *preceding* it. So, in the sentence: *"At the end of last year, the government/gatabont prepared a lengthy report on birth control"*, /g/ is the critical, and /p/ the target phoneme. It is assumed that the derivation of a phonetic representation of the input from the acoustical signal can be accomplished equally rapidly for both words and nonwords. Therefore, it is predicted that there will be little or no difference in RTs when the target is carried by a nonsense word (such as *gatabont*) vs. a real word (*government*).

However, position information is required to determine that the target phoneme is word-initial. This could be achieved by (1) accessing the target-bearing word in the lexicon, because word boundary information is inherent in the phonological representation of a word, or (2) by using the knowledge that the preceding phoneme is word-final. When a target phoneme occurs *after* a nonword, the second way of determining whether a target phoneme is word-initial is not available and the first is severely slowed. This means that the word boundary cannot be determined in a "top-down" fashion, i.e., based on knowledge about possible words. This was shown to be reflected in longer RTs after nonwords.

Similarly, when using only *real* words and manipulating word frequency, there is again no difference in the "*On* condition", but significant differences have been found when the target occurs *after* the low- vs. the high-frequency items (shorter RTs after high-frequency words), because of a similar word-boundary problem. This supports a role for a bottom-up procedure in phoneme identification, in which at least part of the identification is carried out directly via the acoustic properties of the stimuli. When a word-initial phoneme that is similar to the specified target is encountered, the "monitoring device" will have a tendency to respond. Therefore, resources will be devoted to this part of the input: a more complete analysis of the critical segment may be instantiated in order to avoid false alarms, so that less attention can be paid to the next part of the sentence (which actually contains the target). As a result, identification of the target phoneme will then have to occur on the basis of the phonological code. Indeed, Dell and Newman (1978) have found that target phonemes preceded by similar critical phonemes are responded to postlexically (reflected by lexical effects in PM RTs), whereas dissimilar critical phonemes lead to a prelexically-based response to the target phoneme.

Within the *dual code* model of Foss and Blank (1980), it is therefore assumed that both phonetic and phonological codes are computed by the speech processor. Therefore, there are two alternative "outlet points" for phonemic decisions: one served by prelexical, the other by postlexical information. Responding to either of these codes requires attention. When some of the segmental and perhaps featural information has been assigned, this representation forms the (partial) basis for lexical access. Prelexical representations (i.e., feature information) fade rapidly and are difficult to access (see also, Fujisaki and Kawashima, 1969,1970; Pisoni, 1973), unless a substantial fraction of processing resources is devoted to a particular task. One interpretation is therefore that the increasing probability of lexical access leads to a higher probability that PM responses are based on the more stable phonological representation than on the more rapidly fading acoustic-phonetic representation. Nevertheless, it is possible to use postlexical information even if it becomes available after a prelexical code has been derived. Which outlet is used is determined by a variety of experimental variables, including attention and processing demands that prevent close monitoring of the rapidly fading prelexical code.

Other models, such as the *Race* model (Cutler and Norris, 1979) consider the prelexical and postlexical routes on which phonemic decisions are based as functionally autonomous, *competing* routes (corresponding to the two components described in the Shortlist

model, see figure 5.3). The response is based on the analysis of either of these routes, where the prelexical representation is considered to be quite robust.[62] Normally, PM decisions will be based on the initially available, prelexical representation. Increasing the likelihood of lexical activation, leads to a higher probability that the PM response is based on a postlexical representation.

However, as has been shown by Eimas et al. (1990), it is not processing demands per se that shift the outlet point from a prelexical to a postlexical locus, but rather a shift in attention. When a secondary task is performed in which listeners are required to examine information that is stored as part of the mental lexicon, listeners probably find it cognitively more economical to maintain a single attentional focus and base their PM responses on a phonological representation instead of basing them on prelexical codes and then shift attention to postlexical codes in order to perform the secondary task. With a short delay between the response requirements for the primary and secondary tasks, processing demands for shifting attention are excessive, so codes on a single level of processing are used that satisfy the demands of both tasks. But when there is sufficient time to shift attention between performing the primary and secondary task (as is the case with for example a 900 ms. delay), listeners seemed to perform their PM decisions based on prelexical information and then shifted attention to performing the secondary task.

Both the dual code model and the race model agree that phonemic decisions can be based on prelexical or postlexical representations, but they differ in whether prelexical codes form the input to lexical items, and whether processing demands or attention shifts explain the increased probability of postlexically based decisions.

Notice that the ability to respond prelexically on PM tasks, in itself does not entail that, during normal speech recognition, information is integrated into phonemic units (see also section 5.3.2.4). This is one of the reasons why prelexical and lexical *processing* are dissociated as functionally autonomous in Shortlist and Race.

5.3.2.2 *Phonetic categorization*

There is another line of research that could be interpreted as providing evidence for the psychological validity of phonemes, namely the results from *phonetic categorization* experiments. Within the phonetic categorization task, subjects are asked to label unambiguous and ambiguous speech sounds, selected from a continuum, ranging from one unambiguous endpoint through an ambiguous region to another unambiguous endpoint. One or more acoustic cues to a phonetic distinction (for example, voice onset time (VOT) in a voicing continuum between /b/ and /p/) are manipulated by speech synthesis or by editing natural speech (for a review, see McQueen, 1996). The task is normally *Two-Alternative* (i.e., the labels for the endpoints of the continuum) *Forced-Choice* (2AFC). Listeners can hear a simple list of experimental items, but they can also appear in phrases or sentences. A categorization decision is required. Plotting the proportion of labeling responses to one alternative often leads to an ogive shape, with a steep slope (responses change rapidly from one alternative to the other) in the center of the continuum. This region is called the *category (or phonetic) boundary* (see figure 5.6).

[62] Note that, dependent on the assumptions underlying the model, the *prelexical representation* of phonemes does not necessarily refer to the acoustic-phonetic input that forms the basis for lexical access. In Race, it is assumed that the phoneme representation derived from acoustic-phonetic information for the purpose of making phonemic decisions differs from the one derived during the process of normal word recognition. Therefore, there are separate routes by which phoneme information can be derived, leading to different codes on which decisions on phoneme identity can be made: one in which lexical processing takes place, and one in which prelexical processing takes place. The prelexical code that is derived by paying more attention to the sound structure of the speech signal (during prelexical processing) is more robust than the one that forms the input to the mental lexicon in the process of lexical access (during lexical processing).

The task has often been run in parallel with a *phonetic discrimination* task where subjects are tested on their ability to distinguish between items on the continuum. Discriminations within a category tend to be more difficult than between categories. So, whereas a similar variation along the manipulated acoustic dimension can be discriminated when the two stimuli fall within different categories, the same difference can be less accurately discriminated between stimuli belonging to the same category. These effects have been known as the phenomenon of *categorical perception*, because subjects act as if the only information available is the phonetic category of the stimulus.

The question remains however, whether these effects are task induced (because listeners have to make two-alternative forced choice labeling responses or because attention is focused on phonetic codes), or whether this is a perceptual effect reflecting a true insensitivity to the acoustic signal properties. In the latter case, the purpose of the categorization of bundles of acoustic features into a higher order phonemic unit would be to abstract away from the details in the signal in order to aid further processing at even higher order units (i.e., syllable, or word units). Recent evidence indicates that the effects very much depends on the stimuli that are used and the task demands, and therefore on the focus of attention. Categorical perception typically occurs when listeners are in the phonetic mode, performing a 2AFC task in which they have to label phonemes (Gerrits, 2000).

The *motor theory* of speech perception (Liberman, 1967) proposes that the feature information that gives rise to phoneme identification is inaccessible, and that this explains the pattern of results obtained in phonetic categorization tasks. Within the dual code model, it is assumed that feature information decays rapidly. This is especially true for consonant features for which weak physical, and therefore neural, evidence exists (due to their short duration and low energy). Therefore, though responses can be based on precategorical (featural) information, responses must often be based solely on the output of the phoneme identification process. Thus, a loss of information from the feature level is coupled with a reliance on a more abstract code.

Within the computational model Trace (McClelland and Elman, 1986) phonetic units are incorporated as intervening between feature and word units. The Trace model can simulate the phonetic categorization results, and is also consistent with a huge amount of other experimental data. Though McClelland and Elman mention that the choice for the representational units is not truly essential for demonstrating their main interest, namely the potentials of their highly interactive processing architecture, it is at least interesting that their specific choice for phonemic units seems to be consistent with so many data. The Trace account for the phonetic categorization data is that the categorical output is the result of an interactive competition process, which is accomplished through inhibitory connections between the phoneme units. This greatly sharpens the differences in the activation of the detectors for the relevant units, thereby explaining the qualitative shift in listener's identification responses.

Furthermore, Trace can also account for the discrimination results obtained within the phonetic categorization paradigm. Whereas the competition process is responsible for the fact that one phoneme gradually comes to dominate the others, the excitatory feedback from the phoneme level to the feature level imposes a canonical pattern of activation on the feature level. As a result, the original pattern of activation is gradually replaced and the remaining pattern of activation assimilates further to the prototype as time passes by. In this way, feature patterns for different inputs assigned to the same category are rendered nearly indistinguishable. This tendency towards categorical perception is combined with a flexibility in feature interpretation, i.e., Trace can account for the context-sensitivity of the variant cues to phoneme identity, and has the ability for trading relations between features. Within the featural domain, one phoneme is signaled by a number of different cues. Human subjects can trade these cues off, which is useful because the trading relations are related to co-articulation effects in speech production.

Such trading relations have been shown in experiments performed by Summerfield and Haggard (1977). For example, the boundary between /ga/ and /ka/ shifts to longer VOTs when F1OF (the onset frequency of the first formant) starts off lower rather than higher and vice versa. Also, in a classic trading relations experiment by Denes (1955), there are clear cases in which a cue that favors one of two phonemes to a moderate degree will give rise to the perception of the other phoneme when paired up with a strong cue that favors the other phoneme. In Trace, such relations are naturally incorporated as the net bottom-up input to phoneme units is just the sum of all of the inputs. No one input is necessarily decisive. How strongly the input will activate a particular phoneme unit relies on the weighted sum of the input. Thus, in spite of the fact that Trace is quite flexible in the way it combines information from different features to determine the identity of a phoneme, the model is also quite categorical in its overt response. This results from the integration of features into phonemes, and the subsequent lateral inhibition between phonemic units at the phonemic layer.

Also, lexical context effects have been found, in the sense that listeners tend to be biased to categorize stimuli in such a way that they form words, but only in the ambiguous region, or in noisy conditions. This effect could be a response bias effect, a true perceptual bias effect (criterion-shift), or a change in sensitivity of perceptual processors through interaction between higher levels and lower levels of processing via "top-down" feedback (see section 5.6). Other effects are for example:

- *selective adaptation effects*: a shift in labeling responses as a result of repeated exposure; also for phonemes sharing phonetic features with the to-be-labeled phonemes (see section 5.3.1.2),
- *co-articulation effects*: "compensating" labeling responses for preceding context (see section 5.6.4),
- *phonotactic (il)legality effects*: biased responses toward phonotactically legal sequence of preceding and/or following phonemes (see section 5.6.3).

It is useful to compare the interpretations made within the localist connectionist framework of Trace with alternative interpretation of the effects that are found. It was mentioned in section 2.2 that phonemes are defined as the minimal units that linguistically contrast words, i.e., the meaning of a word changes when one phoneme changes. Because of the inability to characterize a phoneme such that it also captures its allophones, it was argued that not the phonemes, but their *features* are linguistically contrastive. Furthermore, the nature of allophonic variations could be more easily formalized when referring to the syllable context in which the phoneme and its features occurred. This way, the underlying phonemes can be derived by applying phonological rules within a syllable context.

Since the evaluation of a feature always takes place within the (syllabic) context in which it occurs, it is plausible that the auditory system has encoded these regularities such that the underlying representation captures the temporal patterning of these features. With an accurate representation of these dynamics, it is not needed to compensate for coarticulation effects. For instance, a feature that distinguishes the phonemes /b/ and /p/ is voice onset time (VOT). This is a feature characterizing the *temporal relation between two auditory elements*: the noise burst that signals the release of closure that characterizes stop consonants, and the presence of voicing in the subsequent vowel. Without the presence of this context, there is no means to distinguish between the two consonants.

Because this feature linguistically contrasts the /b/ and /p/ in words (where they are followed by a vowel), it is a behaviorally significant feature, and therefore an information-bearing element (see section 3.9.2). Based on the distributional regularities of the occurrences of /b/ and /p/ in different words, the range around which these stop consonants can

be distinguished by the parameter VOT lies around 35 ms, which corresponds to the phoneme boundary. Therefore, a VOT of 35 ms forms an *information-bearing parameter that allows for the discrimination* between /b/ and /p/ *in vowel contexts*. As a result, the perceptual encoding of this feature *is* more sensitive around this range making it *perceptually more discrete*, but *only when the context is present*. It can therefore be argued that not the phoneme, but the syllable in which it occurs is categorically perceived.

From this point of view, it is not the feature that is re-interpreted after the recognition of the context. Rather, the presence of the context ensures the ability to effectively discriminate on the basis of a feature to which we have learned to become more sensitive *as part of the context*, and therefore only when relevant. A dynamic feature representation encompasses this context. When it is assumed that the dynamics of speech can be accurately represented as a result of frequent language exposure, the regularities that characterize allophonic variations are directly encoded. There is no need to compensate for the absence of a feature by first "recognizing" the context and then reconstructing the feature values such that it matches the presumed prototype description (which is never realized!). The same reasoning holds for the application of phonological rules to derive an underlying phonemic representation. Such rules are useful for descriptive purposes, but do not characterize the processing that takes place, since they are also implicitly encoded.

5.3.2.3 *Articulation scores*

Taking a completely different approach, Allen (1994) argues for a completely hierarchical system (with no feedback) where frequency-local speech features (e.g., formants) are integrated into sound units (phones), which in turn are integrated into syllables, words, sentences and meaning. It seems that humans work with partial recognition information across frequency for which independent feature-processing channels are required. For example, it has been shown that partial recognition errors across frequency are independent. Forcing such errors in one frequency region does not affect the partial recognition at other frequencies.[63] Allen states that this makes human speech recognition so robust. All that matters is the *local* SNR within a channel, and *the role of the formant is to change the SNR* (Allen, 2000).

Allen (1994) also refers to studies performed by Fletcher and colleagues at Bell Labs between the years of 1918 and 1950. Effects of filtering and noise on speech recognition accuracy were studied for nonsense consonant-vowel-consonant (CVC) syllables, words, and sentences. Using nonsense CVCs removes the effects of context - that can be used across levels - which are present when using word stimuli, or sentences. As words are the subsets of these CVCs that have meaning, the entropy of words is lower relative to nonsense syllables. Due to the meaning conveyed by a word or sentence, listeners can compensate for missing phone information.

Syllable articulation $S(a)$[64] can be accurately predicted from the phone articulations $c(a)$ and $v(a)$ by the relation $S(a) = c^2v$, which reflects that the individual sound units are independent. This seems to imply that coarticulations of the speech sounds are transformed by the auditory system into independent units at an early stage, before context is used, since context was not present during testing. Allen argues that the interdependence between dif-

[63] This property does not hold for current ASR systems. When using spectral templates, the errors across frequency are not independent. As a result, when presented with noise, filtering, reverberation, multiple speakers, and other degradations, ASR systems are not robust.

[64] *Articulation* refers to the empirical probability of correct recognition when context is *not* present (e.g. recognition of nonsense syllables). *Intelligibility* refers to the empirical probability of correct recognition when context *is* present (e.g., recognition of words). The α in the formula indicates that the articulation and intelligibility scores depend on the local SNR within an articulation band or frequency channel. When $\alpha = 1$, the conditions are ideal, and when $\alpha = 0$, the SNR is very poor.

ferent speech sounds as a result of production, is irrelevant for perception. Whereas many speech researchers believe that the coarticulation makes the speech recognition problem difficult, according to Allen (1996), this is all based on the misconception of trying to find a one-to-one correspondence between spectral templates and phones. He concludes from the syllable articulation experiments that phones are perceived independently.

Furthermore, independent sets of phone features (implying the recognition of *partial information*) are analyzed within independent articulation bands[65] (frequency channels) and processed independently, up to the point where they are fused to produce the phone estimates. The phone features are assigned in a categorical manner to the phone class based on local binary decisions regarding the presence or absence of the features[66] (comparable with the binary distinctive features described in 2.2.1.4). Allen (2000) states that psychophysics is needed to gain insight of how features are extracted and what the physical parameters supporting each feature are.

It should be noted though, that the use of nonsense syllables does not only remove the influence of meaning that is conveyed within words. Using such atypical stimuli that do not comprise part of our language also changes the adequacy of using a learned (prelexical) sensitivity to phonological regularities that characterize a language, and therefore normal language behavior. Within the process of speech recognition, it is not the *presence* of meaning that is conveyed within a word that changes the interpretation of the phoneme. Rather, the fact that a word's meaning is *distinguished* by changing one distinctive feature, directly influences the interpretation of this feature as part of the word. It can therefore not be guaranteed that the task of recognizing nonsense CVCs taps into the process of normal speech recognition. It is very likely that the nature of the used stimuli changes the way the speech sounds are processed. Therefore, the task is actually not to identify nonsense syllables, but to attend to the individual phonemes that make up these syllables. The independence that is found between the accuracy of recognizing the syllables and the accuracy of recognizing its constituting phonemes might therefore be task-induced. Nevertheless, it does reflect part of our language ability to attend to the sound structure of our language corresponding to individual phonemes. However, as will be seen in the next section, this ability does not seem to be part of normal language behavior where the purpose is to construct a mental model of the scene by interpreting the meaning that is conveyed by the relations between different words within a sentence context (or in a broader discourse context), and by using background knowledge.

---

[65] The partial articulations (for each band) did not sum to the wide band articulation, but a nonlinear transformation did make them additive. As we are dealing with probabilities, this additivity condition is basically an independence argument.

[66] Notice the difference between the interpretation of co-articulation effects and categorical perception when comparing Allen's reasoning with accounts such as the Trace model (or the FLMP model). Both assign features to phonemes through an integrative process. But, in Trace, the features are not binary valued, but have *activation values* that are related to their reliability (which is of course also related to local SNR values as it depends on the *amount* of evidence). Furthermore, there are trading relations between feature values, which implies that they are not completely independent in their *interpretation*: acceptance of a feature as belonging to a particular phoneme *depends* on the values on other phonemes. Though it does independently contribute its activation to the sum of the activation value for the phoneme, it does not become a feature (i.e., an interpreted cue) unless the combination of feature values is consistent with the overall phoneme interpretation. It seems that coarticulatory aspects influence this interpretation. Allen states that the assignments are based on *local binary decisions*, and he actually argues for the irrelevance of coarticulatory consistency, despite the fact that in speech production the realization of a feature naturally depends on other features, and the preceding and following context.

5.3.2.4 *Speech as a sequence of phones?*

Despite the usefulness of the concept of phonemes (or phones) for descriptive purposes, and the intuitive appeal to consider the lexical representation of spoken words as a string of phonemic labels (essentially an auditory analogue of the representation of written words as strings of letters), care should be taken in the interpretation of results from phonetic decision tasks, and other behavioral tasks such as identifying nonsense syllables or words.

The ability to deal explicitly with the phonetic units of speech is not acquired spontaneously. Research has indicated that there is a relation between our awareness of speech as a sequence of phones and our ability to read (Morais et al., 1979; Bradley and Briant, 1993). For example, speakers of languages where individuals learn to read purely logographic and thus have no training in phoneme-grapheme correspondences have difficulties with phonetic decision tasks (Cutler and Otake, 1994). In addition, Morais and colleagues (1979) showed that adult illiterates cannot add or delete a phone at the beginning of a nonword. Their performance on such tasks is extremely poor compared to the performance of adult literates, and even slightly inferior to the performance of children at the age of 6 that are just learning to read. The fact that illiterates are not aware of the phonetic structure of speech does not imply that they do not use segmenting routines at this level when they listen to speech. But there is a risk in studying the mechanisms of speech perception through tasks that require conscious, *explicit* segmentation. If the question concerns how we perceive speech, by first segmenting it either in phones (phonemes) or in syllables, then it refers to *implicit* (tacit) knowledge. Therefore, it is important to distinguish between the prevalence of such or such a unit in segmenting routines at an unconscious level and the ease of access to the same units at a conscious metalinguistic level.

5.3.3 Feature-based mappings

Marslen-Wilson and Warren (1994) devised a number of experiments in which they tried to distinguish between a phonemic view and a featural view regarding the mapping onto lexical representations. They argue that the input to the recognition lexicon must be specified in featural terms. *Speech is produced as a continuous sequence of articulatory gestures, which results in a continuous modulation of the signal, which is faithfully tracked by the processes responsible for lexical access and selection*. Several types of *partial information* about incoming segments are made *immediately available* to the lexical level, without needing to wait until the current or subsequent segment has been identified. For example, after a vowel, available cues are immediately used to restrict the class of possible lexical choices, under conditions where it is still not possible to know the full identity of the postvocalic consonant. They claim that this can be explained by assuming that there is no prelexical integration of featural cues to identify higher order units.

5.3.3.1 *Subcategorical mismatches*

This conclusion is strengthened by the results from tasks with *subcategorical mismatches*, i.e., tasks in which stimuli are used where the initial part of a (non)word is cross-spliced with the end of another (non)word. This brings different cues to phonemic (or category) identity into conflict. Subcategorical mismatch effects are reflected by an increase in error rates and/or increasing RTs.

Type of sequence	Notation	Example
Word sequences		
1. Word 1 + Word 1	W1W1	*job + job*
2. Word 2 + Word 1	W2W1	*jog + job*
3. Nonword 3 + Word 1	N3W1	*jod + job*
Nonword sequences		
1. Nonword 1 + Nonword 1	N1N1	*smob + smob*
2. Word 2 + Nonword 1	W2N1	*smog + smob*
3. Nonword 3 + Nonword 1	N3N1	*smod + smob*

Table 5.3 Examples of the stimuli used by Marslen-Wilson and Warren (1994) to generate subcategorical mismatches.

As can be seen in table 5.3, the stimuli used differ in their properties with respect to the lexical level (in the sense that they make either word or nonword sequences), but provide phonetically parallel mismatches. Earlier research had already indicated an absence of subcategorical mismatch effects for nonwords, and a presence of such an effect for real words (Streeter and Nigro, 1979; Whalen, 1984).

According to Marslen-Wilson and Warren (1994) this is because there is no prelexical level of representation at which these disruptions can be detected and that it is only with respect to a lexical representation that these cross-spliced sound sequences are anomalous. If prelexical processes compute only a featural analysis of the speech signal, then conflicts in successive featural cues to place, as in the mismatches described above, will only have an effect at a level of analysis where these featural cues are *integrated* to form higher order units, which on a featural account of lexical representation and access, takes place at the lexical level and not before. They therefore predict that phonetic mismatches should not disrupt processing under conditions where they do not map onto lexical form representations (as is the case when both parts of a cross-spliced nonword stimulus are taken from nonwords). A phonemic hypothesis prediction, on the other hand, would state that *all* mismatches should disrupt output to the lexical level.

The stimuli were (C)CVC monosyllabic (non)words, that were spliced at vowel offset, so that mismatch is between coarticulatory cues in the vowel and cues in the consonantal release. They used mostly (un)voiced plosives[67] to ensure that pre- and postsplice cues to a given phonetic property were sufficiently strong to conflict with one another.

They first used an *auditory lexical decision* task, where listeners have to make "Yes" or "No" responses regarding the lexical status of the stimuli by pushing a button as soon as possible. The results indicated that the consequences of subcategorical mismatches depended on the lexical status of the conflicting cues, such that conflicts that only involve nonwords (as in the N3N1 condition) do not disrupt performance. For the word stimuli the responses were markedly slower and more errors were made for both mismatch conditions (W2W1, N3W1). For the nonword stimuli only the W2N1 condition led to significantly slower responses, probably because of the initial word activation provided by the W2 part. This pattern of results is consistent with a featural theory of lexical input.

[67] Previous research using the gating paradigm has shown that for *unvoiced* plosives place preferences are already building up at the lexical level before the end of the vowel (Warren & Marslen-Wilson, 1987, 1988) and phonetic research also shows an early interpretation of cues to place (Repp, 1980). Therefore, mismatch effects are expected, though probably weaker than for *voiced* plosives, because research performed by Malecot (1958) and other research by Warren and Marslen-Wilson (1987) suggests that pre-release cues to the identity of postvocalic voiced and unvoiced stops play a more important role in determining place of articulation for voiced than for unvoiced stops. These are therefore the most likely ones to provide stimulus contrasts where the pre-splice cues are strong enough to conflict with post-splice cues in consonantal release.

To test whether the result of the previous experiment really reflected lexical activation effects and did not reflect low-level effects of cross-splicing, a gating task was used. Within the gating paradigm subjects are presented with successively larger fragments of a word until the entire word is heard. On each trial, a subject has to indicate the identity of the presented item. In this case, the presented gates were increased with 25 ms on each trial. The results showed that the overall patterning of uptake of place information was qualitatively very similar for Word and Nonword mismatch stimuli indicating that the phonetic cues to place are equivalent for the critical N3W1 and N3N1 conditions. Therefore, there is nothing in the relative effectiveness of these pre- or postsplice cues that could explain the large mismatch effect for N3W1 in the lexical-decision task and not for N3N1.

However, there were again marked differences in the effectiveness of presplice cues to place as a function of the lexical status of the fragment heard: nonword initial fragments were less effective cues to place. These data suggest that coarticulatory cues to place of consonantal articulation do not lead to strong place hypotheses unless there is a unique lexical item onto which this place information can be mapped and is also compatible with the other constraints on the phonetic class of the relevant consonant. This confirmed that word and nonword initial fragment did differ effectively in their lexical status.

According to the featural account this is explained as follows: As speech is heard, some set of features will be extracted and mapped onto the lexical level. The initial fragment of a word will allow for a complete match of available feature information to a given lexical representation, which is not the case for a nonword initial fragment, because it does not match completely with any lexical representation. But, it does match partially with other real words which will tend to be given as responses.

Finally, a *phonetic decision* task was used in which speeded responses regarding the phoneme identity of the postvocalic consonant had to be made. The idea was that if phonetic cues are used to identify larger units, such as phonemes or syllables, then this process should always be disrupted by the presence of subcategorical mismatches. In this task, listener's attention is explicitly directed toward prelexical processes. The results showed that for the Word stimuli, the mismatch effects were almost equally strong, whereas for the Nonword stimuli there was a much reduced effect for the N3N1 condition compared to the W2N1 condition. In this case the mismatch effect was increased compared to the lexical decision task, but there was a same reduction (if not disappearance) of the effect in the N3N1 condition. This cannot be simply a consequence of using a task that directs listener's attention to the lexical level, which was the case in the lexical decision and gating tasks.

To explain the differences between word and nonword stimuli, Marslen-Wilson and Warren suggest that nonwords are perceived through the lexicon, in an analogical fashion. They assume a *distributed* computational substrate for the process of mapping from featural inputs onto lexical representations (for instance, an attractor network exhibiting settling behavior, where the system has to settle into a stable state before a perceptual product becomes available for postperceptual analysis). The output of this process produces a lexical percept on which we can operate postlexically, where the output when presented with a nonword will partake of the properties of the lexical items with which it overlaps. This could explain the differential effects for N3N1 and W2N1, as well as the overall slower responses for Nonword stimuli: the stabilization process will be slower for nonwords compared to words, because for words there is an existing state into which the system can settle directly and for nonwords the system has to find a new stable state, which is likely to take longer. Furthermore, it is likely that this settling process will interact with the component parts of the nonword sequences: an initial attraction on the state of the network will have to be overcome before the stable state emerges for sequences of the W2N1 type. For N3N1 nonwords, the stable state will not be subject to the same pull. Both will be delayed relative to N1N1 nonwords, because the cross-splicing causes an abnormal combination of inputs to the system, and similarly for the spliced Word stimuli.

The Trace model (McClelland and Elman, 1986), on the other hand, is a *local* connectionist model which means that there exist semantically transparent units,[68] such as the proposed phonemic units. Nonwords are recognized by recognizing the individual phonemes, but because of the bi-directional excitatorial connections between phonemes and units, the interpretation is highly biased in favor of words. This can explain for example the faster RTs for words compared to nonwords in PM tasks, and the lexical bias effect (Ganong effect) that will be considered in section 5.6.1. However, many simulations of the Trace model with the attempt to account for the results found with subcategorical mismatches, indicate that Trace can not cope with the pattern of intimate interaction between featural, phonemic, and lexical aspects of speech analysis, and can not explain the degree of parallelism between lexical and phonetic decisions. This seems to be a direct consequence of the way dependencies between and within processing levels are organized in Trace (i.e., lateral inhibition between competing items within a level, top-down flows of activations between levels, absence of bottom-up inhibition), leading to the wrong balance between effects of new information coming into the system and the pattern of existing activation between and within levels.

In other research independent evidence against lateral inhibition at the lexical level and supporting bottom-up mismatch having direct inhibitory effects has been found (Marslen-Wilson, 1993; Marslen-Wilson et al., 1991; see also section 5.5.4.1). Marslen-Wilson and Warren (1994) therefore conclude that:

- Processing at the lexical level continuously tracks the featural microstructure of variation in the speech signal.
- Prelexical decisions are inextricably linked with events at the lexical level.
- Theories of lexical access where contact between featural analysis and the lexical level is discontinuous can be ruled out, for example by requiring an intervening processing unit to be fully identified before an output can be sent to the next level.[69]
- Theories that claim that there is a fully autonomous processing level intervening between featural analysis and lexical interpretation and that tasks like phonetic decision can tap directly into events at this level are also ruled out.
  This assumption of independent accessibility of a prelexical phonetic processing site is a necessarily presupposition for experimental tasks that study speech processing by asking listeners to make judgments about the phonemic interpretation of a speech signal.[70]

The results are consistent with the class of models in which *featural cues are mapped directly onto the lexicon* (e.g., the Cohort model and Shortlist). There is no intervening layer of phonemic classification. It is assumed that acoustic cues are extracted from the speech signal which are translated into phonetic features. These are directly mapped onto featurally organized lexical representations.

---

[68] Such units should *not* be interpreted as reflecting single neurons, but reflect a group of neurons that represent the different temporal patterns associated with features, phonemes and words.

[69] However, it should be mentioned that although Trace has an intervening layer of phonemes, processing at *each* level is *continuously* updated in the light of new information. Therefore, Trace allows for *parallel* transmission of acoustic-phonetic features, and does not require a full identification of phonemes before they can have their effect on a lexical level. Though it seems that Trace cannot account for the mismatch effects, it might be that this results from the complex interactive processing architecture of Trace. It is not likely that this can exclusively be attributed to the presence of a phonemic layer based on the simulation results, but the experimental results do seem to imply this.

[70] It is only a necessarily presupposition if it is assumed that having access to phonetic units for making phonetic decisions taps into processes of word recognition implying that these phonetic units are really computed during word recognition. But this assumption does not have to be made, because it is also possible to follow different routes in order to make phonetic decisions, one in which the phonetic identity of a segment in the acoustic signal is computed directly and one in which it is indirectly derived postlexically without assuming that the former process intervenes in the process of lexical access of the latter (consistent with the dual code model, Race and Merge).

Furthermore, there is *competition between multiple candidates* which are activated by the information in the speech signal, and recognition is based on the emergence of the best fitting candidate. But, the system is also highly sensitive to feature deviations between input and lexical representations: *bottom-up mismatch has direct inhibitory effects*. Competition effects (by means of lateral inhibition between competing lexical candi-dates) are mediated at the decision-stage level. By assuming, for example, a distributed computational substrate for lexical processing, a uniform account of the perception of words and nonwords is also possible.

5.3.4 Syllables as units of representation

The ability to continuously track the temporal evolution of features in relation to feature-based lexical representations described in the previous section, actually reflects the temporal dynamics that underlie *both* the production and acoustics of speech. To account for co-articulation effects, advocates of the motor theory of speech perception (e.g., Liberman et al., 1967) suggested that because of the daunting challenge to encode all of the possible coarticulatory effects imposed on formant patterns, it would be more parsimonious to have an alternative representation of speech. The representation they suggested was in terms of the underlying articulatory gestures generating the acoustical signal. This appeals to the sensible intuition that similar mechanisms are likely to govern both the production and perception of speech and that some unifying representation must exist. This could be achieved by a pattern of resonance between the representation that results from speech perception, and the internally generated representation that represents the neural codes to produce the same signal. However, many researches believe that this interpretation of back-tracing the articulatory gestures from the acoustic signal in real-time, and the nature of the needed neurological mechanisms, is biologically far-fetched.

Another possibility is that the temporal dynamics underlying both the production and acoustics of speech is recoverable given an appropriate auditory representation of the acoustic signal (Greenberg, 1996). The ability to create a functional equivalence across many diverse instances of the same basic meaning, which reflects the perceptual invariance of speech, is a key issue for any theory of speech perception. When a stable representation across the full range of acoustic conditions typifying speech can be provided, this renders the speech signal relatively impervious to the potentially deleterious effects of reverberation, background noise and speaker variation. Greenberg proposes that the auditory system captures the temporally dynamic properties of speech through computation of the low-frequency portion of the modulation spectrum, which is remarkably stable over a wide range of acoustic environmental conditions. Such a low-frequency modulation spectrum emphasizes on temporal intervals between 100 and 300 ms, and is therefore suited to extract syllabic and related phonetic information required for accessing higher-level linguistic representations of the speech signal. Greenberg's proposal is based on an integrated account for the perception of speech and his arguments will be considered in more detail in the rest of this section.

5.3.4.1 Against the importance of spectral detail for representing speech

Traditionally, the auditory system's primary function was seen as computing a running spectrum of the acoustic signal for subsequent conversion into linguistic units by other parts of the brain (Klatt, 1989). So, the role of the auditory pathway was confined to feeding spectra into a phonological processor that churns phone sequences into lexical units. Such a *passive*, bottom-up view of speech processing, however, can not account for phonemic restoration effects, where complete occlusion of an entire phonetic constituent by an interfering sound is hardly noticed and has no effect on intelligibility (e.g., Warren, 1982, see also section

5.6.5). The same argument holds for periodic and random interruptions and deletions of speech. These also have no significant impact on the decoding process (Miller and Licklider, 1950). These effects demonstrate the redundancy of speech where phonetic features are signaled by many different cues, distributed in both time and frequency. However, at that time the specific nature of the representational distribution remained unspecified.

Furthermore, the spectro-temporal constraints underlying the perceptual organization of speech and auditory function (Bregman, 1990; Darwin and Carlyon, 1995; McAdams, 1993) can not be readily integrated into a unified theoretical framework for understanding speech perception (see also chapter 4). However, in some computational ASA models these ideas are more refined, and it has been shown that *complex auditory phenomena can arise from the activity patterns of so-called neural oscillators*. These provide a more concrete, and more biological foundation for the extraction of perceptually relevant properties such as pitch, loudness and timbre (e.g., Lyon, 1984; Weintraub, 1985).

Also, in ASR it has been discovered that spectral characterizations of frames, based on FFTs, are too detailed for adequate generalization of frames of speech to the composite phonetic representations. *Linear predictive coding* (LPC), in which the transfer function of the vocal tract is inferred, improves the generalization. This is consistent with the articulatory perspective on speech perception. Auditory-inspired techniques as *Mel-cepstrum* and *Perceptual Linear Prediction* (PLP), have been shown to provide even better generalizations. With these techniques, the portion of the spectrum below 1500 Hz is emphasized, the spatial frequency organization is commensurate with that of the human auditory system, and a highly smoothed spectral envelope is formed. Furthermore, PLP incorporates additional information that is more consistent with the internal auditory representation of static spectra (such as spectral integration and a compressive loudness growth function). Another form of spectral conditioning, *RASTA*, emphasizes the dynamic components (about 4-16 Hz) of the speech signal, and has also been shown to improve ASR performance under certain noisy conditions. However, there is still a great disparity between the performance of ASR system under ideal (relatively noise-free acoustic environments) and more realistic conditions.

Greenberg also notices that the relative sharp tuning of AN neurons (within the range of 30-40 dB SPL) would appear to provide an abundance of spectral detail. However, under the relatively high sound levels under which speech is generally produced (40-60 dB), the tuning broadens appreciably, *particularly when measured in terms of neural synchrony*. This implies that typically only a few spectral features of the signal, associated with spectral peaks, are actually encoded in the peripheral discharge patterns. This may relate to the effectiveness of the dimensionality reduction in the representation of speech spectra when using techniques such as PLP.

Also, because of the large range of acoustic variability, it seems that a faithful, detailed representation of the speech spectrum would actually serve to impede effective generalization across the normal range of acoustic variation encountered. Therefore, an effective means to deal with this overload of spectro-temporal detail could be to encode only a sparse representation of the speech signal encapsulating only the linguistically relevant information. This is strengthened by evidence from cochlear implant patients who can achieve a remarkable high degree of performance benefit from crude electrical stimulation patterns that provide little more than the low frequency modulation patterns distributed over just a few spectral channels. Also, from evaluations with channel vocoders, it seems that, as long as the bands partition the spectrum in a manner consistent with the range of lower formants (Band 1 < 800 Hz, 800 Hz < Band 2 < 2500 Hz, Band 3 > 2500 Hz), only three or four channels are required for effective intelligibility.

5.3.4.2 *Acoustic shielding of perceptually relevant features*

It is thought that neural phase-locking plays an important role in the spectral processing of complex signals. In chapter 3 it was seen that this phase-locking can accurately track the temporal fine structure of a signal. So, if a detailed spectral representation is not required for adequate intelligibility, then what might be the role of this important medium of neural encoding for the processing of speech?

A likely function is to shield the informational constituents of speech from the deleterious effects of (aperiodic) background noise. Many auditory neurons preferentially discharge to low-frequency, quasi-periodic components of the acoustical signal. This synchronized activity provides an effective means for the informationally relevant components of the signal to "rise above" the background. This is not the case for noisy signal components. By virtue of their statistical properties, these are not nearly as effective in capturing the temporal properties of such neurons (Greenberg, 1988; see also section 4.12.1).

Interestingly, when looking at the spectro-temporal properties of speech, there is a nearly continuous presence of voicing (i.e., pitch), a predominance of energy in the low-frequency (< 2.5 kHz) portion of the spectrum where neural phase-locking is strongest, and a prevalence for abrupt onsets for syllabic components. These properties therefore provide a robust medium for transmitting information under variable acoustic conditions.

This suggests that a primary selecting factor shaping the acoustics of speech was the ability to withstand the deleterious effects of the acoustic background (Greenberg, 1995) and to provide a means of grouping together the neural activity evoked by related spectral elements (Cooke, 1993; Shamma, 2000). For instance, the fact that all harmonics of a certain F0 fire synchronously with the period of the fundamental enables them to become perceptually more distinct *as a group*. The grouping of harmonics can therefore be considered as an emergent property that results primarily from the temporal properties of groups of neurons.

It seems that the spectro-temporal properties of speech are not only constrained by the human vocal tract apparatus that sets limits on the range of realizable spectral configurations, but is also determined by the constraints that are imposed by the auditory system. Relevant parameters of auditory functioning are:

- the range of frequency selectivity
- the frequency resolving capabilities of peripheral auditory neurons
- the limits of neural periodicity coding
- the time course of rapid adaptation
- the temporal limits of neural coincidence detection
- the modulation transfer characteristics of brainstem and cortical auditory neurons

Greenberg (1995) argues that these parameters can account for the following important acoustic properties of speech:

- an absence of spectral energy above 10 kHz
- a concentration of spectral information below 2.5 kHz
- a preference for encoding perceptually relevant information in the spectral peaks
- sound pressure levels for segments ranging between 40 an 75 dB
- rapidly changing spectra
- a prevalence of phonetic segments with abrupt onsets and transients
- a preference for quasi-periodic waveforms whose fundamental frequencies ranges between 75 and 330 Hz
- temporal intervals for integrative units ranging between 50 and 250 ms such as syllables and (phonetically-related) segments

5.3.4.3 *Multiple time-scales*

Another interesting fact is that listeners appear capable of combining information *across* spectral regions. The outcome of this integration can far exceed the sum of the analyses performed separately on the constituent bands. This suggests that inferential tracking of the temporal dynamics over a few spectral regions is involved in the process of speech decoding. A detailed spectral representation provides the means to do this, but according to Greenberg, it is neither necessarily, nor in certain circumstances, desirable. The mechanisms underlying this cross-channel integration are not well understood, but the time dimension is certainly involved.

Multiple time-scales have been revealed by shifting channel outputs in time relative to one another. It has been shown that speech can withstand channel decorrelations as long as circa 120 ms without significant loss in intelligibility. This indicates the existence of at least two separate levels of analysis: (1) one based on local, probably within critical-band information; such local analysis of the spectrum appears to occur within milliseconds, and (2) one based on global integration across channels; spectrally more global analyses often require tens to hundreds milliseconds to perform.

In accordance with the notion that intelligibility depends on the integrity of the low-frequency modulation spectrum, it has been shown that the intelligibility of interrupted speech mentioned earlier depends on two parameters: (1) a minimum acoustic duration of circa 40 ms for individual segments, and (2) an interval of successive segments of not more than 200 ms (Huggins, 1975).

Interestingly, these crucial time intervals of 40 and 200 ms correspond to event rates in the auditory nervous system. Neural discharge rate at the level of the AN and most auditory brainstem nuclei are in the order of 150-250 Hz. Also, at the thalamic level and in the medial geniculate body, discharge rates of 100-200 spikes per second are not uncommon. In the cortex, however, discharge rates rarely exceed 30 spikes per second, and more typically occur at rates between 5 and 20 spikes per second (see also chapter 3). Therefore, the auditory cortex is likely to function as a highly inertial system in which thalamic input plays a relatively subordinate role except to signal major *changes* in the afferent input, and therefore to provide *information*.

The reduction in discharge rate effectively down-samples the auditory representation, and as a result facilitates generalization across diverse instances of the "same" thing. The intra-cortical input likely contains information about the state of both adjacent and distant tonotopically organized elements, as well as about the temporal evolution of the spectrum. The across-channel analysis could occur in at least two stages:

(1) one associated with approximately 40 ms intervals for phonetic, subfeature analysis spanning several contiguous channels, and
(2) a longer interval (ca. 200 ms, roughly corresponding to syllable-sized units) required for integration of subfeatural information into a coherent representation for higher-level linguistic processing.

These longer units are distinguishable on the basis of the composition and order of the shorter subfeatural elements. Hence, *phones* can be thought of as a constellation of subfeatures which, *when bound together across time*, serve as the carrier of linguistic information *through the action of the syllable.*

The 40 and 200 ms intervals correspond to various other aspects of cortical processing, and are not specific for the auditory cortex. For audition, the 200 ms interval corresponds to the length of the temporal integration window (section 4.5.4), and the upper limit for the continuity effect. Furthermore, the typical length of a syllable in fluent speech is circa 200 ms on average. This is sufficiently long as to provide some measure of perceptual stabil-

ity, and short enough to provide a sufficiently dynamic representation of the external stimulation as to be functionally effective.

The 40 ms quantum interval allows the elements within the 200 ms to be distinguishable. It is the minimum segmental interval required for the occurrence of the continuity illusion, and it is about the shortest span of segmented speech that can reliably be associated with a specific phonetic quality. The 40 ms frames into which syllabic units appear to be quantized allows that phonetic distinctions can be made among syllables. Syllables that are distinguishable on the basis of voice-onset-time (VOT) also have their category boundary around the interval of 40 ms. For instance, the perceptual distinction between /ba/ and /pa/ is largely based on VOT, and the perceptual boundary is usually found around 35 ms in humans (Eggermont, 2000). Auditory cortical recordings to such syllables are consistent with this idea.

In addition, the 40 ms interval corresponds to the interval in which acoustic stimulation begins to assume an *independent identity*. For example, when a spectral component begins less than 40 ms prior to the beginning of a vocalic segment, it is not separately perceived from the other components. When the same segment leads by more than 40 ms it is heard as a separate stimulus distinct from the vowel, and with a clearly defined pitch. Nevertheless, its presence still contributes to the vocalic identity (Darwin and Sutherland, 1984). Such a dissociation between the perceived coherent phonetic identity and the separation of sound sources, was also seen in section 4.10 regarding the phenomenon of duplex perception of sound.

5.3.4.4 *The importance of segmentation*

An important role played by the auditory system is to provide segmental information. In general, there is a gain in speech intelligibility, especially under noisy conditions, when visual information pertaining to the articulatory movements is provided in addition to the acoustic signal. This visual analogue provides important information regarding segmentation and the phonetic identity of the syllabic constituents. This can be interpreted as the result of providing an independent source of information concerning speech segmentation and phonetic boundaries by virtue of a temporal dynamic common to both acoustic and articulatory representations of speech. For instance, the hearing impaired often gain significant benefit from speech reading. Moreover, there are some interesting paradoxes of hearing impairment:

- The locus of energy in the speech signal (< 2 kHz) is considerably below the region showing the greatest deficit in sensitivity (typically > 3 kHz). This is inconsistent with the idea that the primary basis for the functional impairment caused by a sensori-neural hearing loss is damage to the spectral analytic capability of the auditory system.
- In quiet, the single best predictor of speech intelligibility performance, is the pure tone threshold for frequencies *below* 2 kHz. In noise, the best predictor of speech intelligibility is the audiometric threshold *above* 2 kHz (Smoorenburg, 1992). This seems to imply that the mid and high-frequency regions are especially important for processing speech under noisy conditions, which is otherwise not so apparent.

It has been shown by Grant and Walden (1996) that the spectral region above 3 kHz is particularly important for delineating the segmentation and number of syllables in spoken language. For understanding spoken language, reliable information regarding syllable *segmentation* appears to be essential (Lehiste, 1970) and to infer the lexical identity of ambiguous speech, syllabic *structure* provides an important means (Segui et al., 1990). Thus, when the low-frequency portion of the spectrum is compromised by background noise, the higher-

frequency portions of the speech signal assume additional significance because of the segmental information that is associated with these channels. It seems therefore, that the ability to understand speech relies as much on segmental analysis as it does on an analysis of the spectrum. In the absence of such segmentation, the ability to understand speech is severely compromised (Drulman et al., 1994). In its presence, comprehension occurs, even with minimal spectral cues (Shannon, 1995).

5.3.4.5 *The significance of the syllable*

Finally, Greenberg (1996) argues against the traditional view that speech can be analyzed as a series of phones concatenated through coarticulation rules imposed by the biomechanical constraints of the vocal tract. He therefore addresses the following question: What if the phone is actually a secondary unit of analysis whose major function is to *distinguish* among different forms of the basic perceptual/representational unit? Greenberg proposes that the *syllable* forms the basic unit of speech perception. This proposal is also based on the following evidence and arguments:

- PM RTs for target syllables are faster than for their constituent phones, even when the target phone is located at the beginning of the syllable (Segui et al, 1990).
- Most effects of consonantal context on vocalic identity seem to be the consequence of intra-syllabic segmentation. When most of a syllable is presented, vocalic identification is high. When significant portions of the syllable are missing, intelligibility is much lower.
- Most co-articulation effects occur within a syllable. Trans-syllabic co-articulation effects are comparatively small.
- Increases in speaking rate result in the deletion and mutation of most phonetic constituents, whereas syllabic units are generally preserved.
- Articulations are generally programmed in syllabic, not phonemic units. Speech-error mispronunciations and "tip-of-the-tongue" recalls are organized on the basis of syllabic, not phonemic units (Fromkin, 1973).
- Integration of visual cues in speech perception occurs over syllabic, not phonemic intervals (Massaro & Cohen, 1993).
- Prosodic properties such as pitch, accent and stress can be more readily incorporated in a syllabic based organization than on the phonemic level.

5.3.4.6 *The modulation spectrum as a representation for speech*

Whereas traditional spectrographic displays undergo dramatic degradation in the presence of background noise and reverberation, the modulation spectrograms proposed by Greenberg are claimed to be remarkably stable under the same conditions. The modulation spectrogram is based on evidence regarding the importance of low-frequency modulation.

According to Plomp and colleagues, phonetic information can be encoded in terms of the slow energy fluctuations that occur across tonotopically organized auditory channels. Long-term analysis of the energy fluctuations indicates a peak in the spectrum at around 4 Hz (Houtgast and Steeneken, 1985), again corresponding to the rate of syllabic units. The necessarily sparse granularity of the spectrum would be sufficient to adequately distinguish among the possible set of phonetic elements (possibly only, or much more accurately, within the context of a syllable).

Speech intelligibility depends on the integrity of the low-frequency portion of the modulation spectrum. Low-pass filtering the modulation spectrum, so that energy fluctuations above 3-4 Hz are significantly attenuated, reduces the intelligibility of speech. This has

been found for Dutch words and sentences (Drullman et al., 1994), and has been replicated for both English and Japanese (Arai et al., 1996). This form of filtering leaves the quasi-steady-state spectral regions relatively intact, but essentially blurs the syllabic boundaries.

Finally, encoding the magnitude of energy in terms of low-frequency modulation leads to a quite stable representation, even under conditions which are known to preserve intelligibility, but disrupt more traditional representations based on spectrographic analysis (Kingsbury, Morgan, and Greenberg, 1988). This modulation spectrogram encodes the magnitude of energy in the lowest modulation band (with a peak at 4 Hz and 10 dB down at 8 Hz, similar to the long-term modulation spectrum of speech) as a function of frequency (quantized by critical band like units) and time (using 250 ms window, and 12.5 ms steps to capture the dynamic aspects). The stability suggests that this information may be sufficient to encode speech-relevant information (Greenberg, 1996). Interestingly, these modulation spectrograms appear to be similar, in many respects, to the population response of auditory cortical neurons, which are most responsive to modulation frequencies below 20 Hz (Schreiner and Urbas, 1988). It may therefore approximate the representation of speech signals in at least some portions of cells in the human auditory cortex.

5.3.4.7 *Statistics of written and spoken language*

The idea that lexical and semantic access requires a complex and rather arbitrary series of operations that proceeds from phonetic units to words, and from words to meaning via a language's grammar does not do justice to what is reflected in the statistical properties of speech. These properties indicate that the relation between the lexical and grammatical elements is anything but arbitrary. Therefore, an overview of such statistically-based acoustic-linguistic associations will be given:

- Only a quarter of the possible words in the English lexicon are one syllable in length, but over 82% of the words in spoken discourse are monosyllabic (compared to 63% in written language).
- For written language the length of a word is inversely correlated with its frequency, which can also be interpreted in informational terms. This probably accounts for the preponderance of short, monosyllabic words in spoken discourse.
- Moreover, RTs for (visually presented) word recognition is also inversely proportional to its frequency of occurrence. Therefore, there seems to be a direct relationship between a word's information content and its duration: the higher the information content of a lexical item, i.e., the less predictable it is, the longer its duration is likely to be.[71] If it is assumed that the temporal properties of language are tailored to synchronize lexical retrieval with the time course of individual verbal elements, it is to be expected that high-frequency words tend to be short (because of their predictability and the relatively short time to retrieve their meanings), while rare words often contain many syllables as they require longer retrieval time. Over the course of a word's linguistic evolution, it will tend to shorten as its usage increases with familiarity (e.g., *automobile* > *auto, car*; *airplane* > *plane*; *refrigerator* > *fridge*).

[71] Though it may be argued that the perceptual system should also be prepared to respond rapidly to *unexpected* stimuli, these statistics reflect that the statistics of the environment are conveyed in the way perceptual *processing* takes place. This makes sense. High-frequency words (by virtue of their frequent usage) represent important information. Therefore, this "expectancy" is depicted in a *bias* in processing in favor of high-frequency words. In addition, language *usage* is also adapted to anticipate this property. Moreover, because of their frequent usage, high-frequency words have more stable internal representations, are more efficiently encoded which allows faster retrieval, and less evidence is needed (which makes them also more noise-robust). This in turn validates their shortening as a result of increasing familiarity, etc. Therefore, there is a close interaction between the environment and the system dealing with the environment.

- Of the 100 most frequently spoken words, all but three are one syllable in length, and most of these are function words, such as articles (*a, the*), prepositions (*of, in*), conjunctions (*and, or*), pronouns (*I, you*) and auxiliary verbs (*have, would*).
- Poly-syllabic words (especially with three or more syllables) tend to be nouns or nominal modifiers, such as adjectives, as if the real-time demands of speech production and perception impede the use of many "high-cost" words. For written language more time is afforded, which allows the writer to use more elegant and specific words.
- Most contemporary writing systems are based on alphabetic (i.e., phone based) systems for reasons of efficiency and economy, not for accuracy of reproducing spoken words. The earliest logographic languages were based on word or syllable units. The current Chinese writing system is essentially logographic-syllabic in format (as Chinese lexemes are monosyllabic). Finally, the mismatch is also reflected in that syllable-based intermediaries have been demonstrated to serve as effective pedagogical tools for children who have not been able to read using the traditional phonics approach.
- The dismissal of the syllable as a candidate for representing lexical information is by virtue of its potentially complex and diverse nature (e.g., CV, CVC, VC in Japanese and Spanish, but in English CCCVCCC (as in *strength*) and CCVCC (as in *bracked*) can also occur). In *spoken* discourse, over 80% of the syllables are of the canonical CV, CVC, VC, V form. Many of the remainder reduce to this format through assimilation and reduction. The remaining exceptions tend to be either low-frequency nouns or inflected verbs (e.g., *look* (CVC) vs. *looked, looks* (CVCC)), where a deviation from the "expected" syllabic form provides important information for inferring its lexical and grammatical status. However, many languages have a more transparent structure, and grammatical markings are typically imposed through affixing rather than through syllable complexification.
- Both written and spoken forms of a language rely on a core vocabulary for expression of semantic information. For instance, in written English, 9 words form 25% of word usage, 69 words account for 50%, 732 words account for 75%. The statistics for spoken English reflect an even greater reliance on a core vocabulary.
- There is a similar reliance on a core body of elements at the syllabic level. For written English, 12 syllables account for 25% of syllable usage, 70 syllables constitute 50%, 339 syllables form 75% of syllabic occurrences. Though it might seem that a syllabic representation of English requires a large number of distinct units to cover the lexical inventory, in actual usage the number of commonly used syllables is relatively small. The remainder can be derived from relatively simple phonetic extensions to the core syllabic inventory. On the other hand, for languages such as Japanese, the total number of separate syllables does not exceed a few hundred.

To conclude, it is proposed by Greenberg (1996) that listeners can understand spoken language by virtue of perceptual strategies that appear to automatically extract syllable-like units through analysis of the low-frequency modulation spectrum. Distinguishing between syllables is possible by means of phonetic distinctions between constellations of subfeatures that are bounded together across time within a time-span that roughly corresponds to that of a syllable. Furthermore, there is a systematic relationship between a language's syllabic structure and its higher level semantics and grammar. Therefore, the brain is able to derive meaning from the speech signal on a continuous basis through a reliance on a core vocabulary of a few hundred, highly familiar lexical items.

5.4 Lexical access

Another dimension on which models of spoken word recognition can be distinguished relates to how they account for the *mapping* between the derived input representation and stored lexical representations. This characterizes the process of lexical access.

5.4.1 Serial versus parallel access

The models can be roughly divided in serial versus parallel access models. An example of a model where a serial comparison takes place is the *search model* of Forster (Forster, 1976, 1989; Bradley and Forster, 1987). Lexical access occurs via a search through a frequency-ordered list of word candidates. Each lexical candidate is compared, in turn, to the currently perceived input, and lexical access occurs when a match is found. This model provides an account of effects of word frequency, whereby high frequency words are accessed more quickly than low frequency words. It is, for example, consistent with the often reported finding that, in noisy circumstances, there seems to exists a bias to recognize high-frequency words at the cost of recognizing low-frequency words.

In contrast, parallel access models propose that multiple lexical candidates are compared simultaneously, with the activation of any given item indicating the current degree of fit of a lexical hypothesis. Here, lexical access is not an all-or-nothing process, but may initially involve the simultaneous partial access of multiple candidates before a single lexical item is recognized. So, a distinction is made between lexical access and recognition. These models can also account for frequency effects, either through postulating differences in the resting activation of the representational units for words with different frequencies (e.g., Morton, 1969), or through stronger connections to units involved in the representation of higher frequency words as a result of more frequent exposure.

Actually, there is fairly broad consensus that the process of lexical access is characterized by multiple, simultaneously active, lexical hypotheses. Therefore, as could be seen in the description of the most important models that try to account for (the dynamics of) spoken word recognition (section 5.2), all current computational models fall within the class of parallel access models. Since the idea of multiple lexical activation has been established within a number of paradigms, some illustrative experimental findings will be discussed in what follows.

For instance, using a *cross-modal semantic priming* paradigm, Marslen-Wilson (1987) showed that multiple candidates where simultaneously activated early in processing. Pairs of words were selected that shared onset information, but deviated before their offset (e.g., *captain* and *captive*). Semantic associates were visually presented prior to the unique specification of an auditory presentation of a word (e.g., the uniqueness point occurs after the /t/ in *capt*), or they were probed at word offset. When semantic associates were presented prior to the uniqueness point, facilitating priming effects (reflected in faster visual lexical decisions to semantic relatives) were found for both visually presented words. These results indicated that both candidates were active. Probing at word offset showed activation only for the word that was actually heard. These findings were replicated (in Dutch) by Zwitserlood (1989).

Within the same paradigm, Connine, Blasko and Wang (1994) showed that in the case of perceptual-lexical ambiguities, again multiple candidates were activated. Lexical items with ambiguous segments (due to the presence of noise) resulted in more than one lexical interpretation. For instance, with an ambiguous onset (e.g., *dent* vs. *tent*), similar facilitating priming effects were found for words that were semantically associated with both words.

Experimental condition	Stimuli: Auditory prime	Stimuli: Visual target	Results*
Visual probing before uniqueness point	*capt-*	*ship* *prison*	Faster visual lexical decision (Yes/No response) for both targets: facilitatory priming effect
Visual probing at word offset	*captain* *captive*	*ship* *prison*	Only facilitatory effect for semantic associate of word that was actually heard
Conclusion: Simultaneous partial activation of multiple lexical candidates early in processing. Activation rapidly reduced when ambiguity resolved (Marslen-Wilson, 1987).			
Unambiguous initial segments	*dent* *tent*	semantic associates	Facilitatory priming effect only for semantic associate of presented auditory prime
Ambiguous initial segments (* reflects the presence of noise)	**ent*	semantic associates of *dent* or *tent*	Facilitatory priming effect for both targets
Conclusion: Multiple lexical candidates activated in case of ambiguity, and still active at word offset. (Connine, Blasko and Wang, 1994)			
Probing at word offset	*trombone*	*rib* *music*	Faster lexical decision for semantic associate of whole word *and* offset-embedded word
Conclusion: Activation of offset-embedded words (Zwittserlood, 1989)			

Table 5.4 Summary of some results found within the cross-modal semantic priming paradigm. (*The descriptions of the results are with reference to the no-prime conditions).

Multiple activation has also been shown for embedded words that are misaligned from onset, as for instance in *bone* and *trombone* (Zwitserlood, 1989). For example, faster lexical decisions were possible for words like *rib*, presented at the offset of *trombone*, suggesting activation of *bone* (see Table 5.4 for a summary of the above described results).

Finally, it has been shown that activation of a word and its "competitors" is not initially influenced by a biasing sentence context. Similar priming effects are found when only the earliest part of a particular word (that is still consistent with multiple candidates) is presented, despite the presence of a semantically biasing context. Context effects emerge only late in a word, and are therefore interpreted as reflecting selection processes (Zwitserlood, 1995), as well as indicating that context does not influence the generation of lexical candidates thereby reflecting the autonomy of generation processes (consistent with Shortlist and the Cohort model).

5.4.2 Ambiguity preservation during processing

The transient activation of multiple lexical items (and associated meanings) is consistent with models that propose that the speech input activates all the candidates that (sufficiently) match the initial spoken input, with selection processes operating to narrow down the set of activated candidates to those that continue to match the speech input. All existing computational models that are trying to account for the dynamics of speech processing have incorporated this idea of multiple hypotheses being simultaneously active (i.e., Shortlist, Trace, Cohort). Such models are called *multiple output models*, because different levels of processing do not produce decisive, single outputs to higher levels of processing. During processing, possible ambiguities are thus being preserved at each level. Of course, such ambiguities are especially present in noisy circumstances, which explains why contextual information typically

exerts a greater influence under such circumstances. This results for example in a *bias* for recognizing words that are consistent with the sentential context as well as biasing effects for high-frequency words compared to low-frequency words, or for words compared to non-words, etc. (see also section 5.6).[72]

5.4.3 Early versus delayed commitment

The impact of the time dependent nature of speech on lexical activation is that as more of the stimulus is heard (over time), the amount of evidence that contributes to lexical activation is greater. However, word recognition could occur earlier when a word can be uniquely identified on the basis of a subset of the potentially relevant information - i.e., when the word passes its uniqueness point - especially for some long words. This would be in line with the basic immediacy assumption and the idea that word boundaries emerge from the recognition of individual words, as incorporated in the Cohort model (Marslen-Wilson & Welsh, 1978). When a word is recognized, the offset of this word specifies where the onset of the following word is. In such a sequential model, recognition and segmentation proceed in strict order. Hence, the correct recognition of a word boundary and therefore the subsequent recognition of the following word seems to depend on successful recognition of the preceding word.

5.4.3.1 Immediate commitments

Evidence in support of these ideas comes from, for example, results of *shadowing* tasks where listeners are aurally presented with sentences and are asked to repeat what they hear, as rapidly as they can. When words are deliberately mispronounced, listeners either give an exact repetition or restore the input to its "intended" form (Marslen-Wilson & Welsh, 1978).

Exact repetitions are typically associated with other disfluencies in shadowing, notably pauses, suggesting that listeners are aware of the mispronunciation and that syntactic and semantic analyses are disrupted by the deviant input. So, subjects heard what was represented in the stimulus, based on a *phonetic* code (see section 5.3.2.1). The associated disfluencies arise either because the appropriate lexical item is also retrieved and the discrepancy is noted, or because higher level analyses are disrupted by the deviant input.
On the other hand, the restorations are typically fluent ones, suggesting that they are true nonperceptions of the mispronunciations, which means that subjects actually heard the *phonological* code associated with the intended word. When variables that increase the likelihood of lexical access are manipulated (e.g., increasing the transitional probability, see also section 5.3.2.1), fluent restorations occur more often (Marslen-Wilson and Welsh, 1978). Also, when mispronunciations occur *late* in a word, the phonological code is more likely to be perceived and deviations will often go undetected. In contrast, when the phonetic code of the *initial* part of the word is mispronounced, lexical access will be impaired and deviations are more likely to be detected. These results are consistent with results from phoneme monitor ing tasks in that syntactic, semantic and intraword constraints influence whether listeners perceive the phonetic or the phonological codes.

[72] At this point, I would like to note that it is not required to explain such effects by incorporating feedback to earlier levels of processing that would lead to an increase in sensitivity in perceptual processing, or to a resolving of the ambiguities that existed at these earlier levels. It is also not necessary to explain such results by postulating that activated hypotheses serve to guide a process that searches for information present in the new acoustic input that is consistent with what is to be expected from previous input. All that is needed is the availability of multiple activated lexical hypotheses that are consistent with the input. Auditory perceptual processes "only" have to deal with the continuous tracking of incoming auditory information, thereby continuously updating the activation of different representations in the light of new input.

Probing of visual target	Stimuli: Auditory prime	Stimuli: Visual target	Results*
After initial syllable	*cap-*	*ship* (related to *captain*)	Small facilitatory priming effect
At word offset	*captain*	*ship*	Stronger facilitatory priming effect
After incomplete prime + delay (equal to missing word portion)	*cap___*	*ship*	Priming effect comparable to intact condition

Conclusion: The impact of a stretch of acoustic-phonetic information is *not static* and does *not* have *immediate* impact on the activation level of a lexical entry (Zwitserlood and Schrievers, 1995).

Table 5.5 Summary of the results of the Zwitserlood and Schrievers study regarding the natural confound of additional processing time and additional acoustic-phonetic information.

Marslen-Wilson (1984) found that fluent restorations of mispronounced words often occurred within 250 ms. It may be though, that the time pressure that is imposed by the shadowing task forces such early decision-making based on less accurate processing. Furthermore, the fluent restorations that occur late in a word might reflect aspects that are involved in speech *production* processes, i.e., the inability to correct (or stop) the execution of a "motor program" corresponding to a particular lexical item when more time has passed by.

5.4.3.2 *The uniqueness point*

Some results on gating tasks have also been taken as evidence for the immediacy assumption. Within the gating paradigm, subjects hear successive presentations of a word, where the first presentation is cut off to only the first *N* ms of the word, and successive presentations are lengthened in *N*-ms increments until the whole word is presented. The *isolation point* is the point at which half the subjects correctly identify the word. It has been found that the isolation point falls very close to the theoretically derived recognition point, which is the uniqueness point. However, the *acceptance point* is the point at which subjects are reasonably sure of the identity of the word, and for this, considerably more input is required.

According to Luce (1986), the concept of a uniqueness point may be limited, because only 39% of common words of the English vocabulary (weighted by frequency) are unique before the last phoneme, 23% of words become unique on the final phoneme, and 38% of words after the last phoneme.[73] This is related to the statistics on the pattern of occurrence of words embedded in polysyllabic words, showing that an overwhelming majority (84%) of polysyllabic words have shorter words embedded within them (based on *syllabic* matches). Positional analyses show that these embeddings are most common at the onsets of the longer word (McQueen et al., 1995).

This seems to imply that word boundaries for these embedded shorter words can not be reliably estimated, and therefore subsequent word recognition would be impaired if it were based on the strictly sequential order proposed in the Cohort model[74] (at least when it is assumed that the activity pattern within a Cohort only sustains during the presence of the bottom-up input). The issue of these embedded words will be dealt with in more detail in section 5.5.3.

[73] Note that these analyses are based on *phonemic* matches. Many words that become unique on the final phoneme could be differentiated earlier based on subphonemic cues related to articulatory processes that are reflected in the preceding phoneme(s).

[74] Again, these analyses are based on matches at a surface level. Other cues might provide sufficient disambiguating information at word offset, for instance subphonemic acoustic cues, syntactic structure, semantic context, etc.

5.4.3.3 *Retroactive effects*

As has been pointed out by Grossberg and Myers (2000), "the brain processes that group sounds into coherent speech units exhibit an exquisite sensitivity to the temporal distribution of spectral energy in the speech stream. This has been revealed in a number of context effects whereby later-occurring information influences an earlier perceptual grouping decision. These so-called *backward effects* directly constrain theories of how the perceptual units of language spontaneously form under variable-rate speaking conditions. In particular, they show that *the time scale of conscious speech is not equal to the time scale of bottom-up processing*.

Though it would seem advantageous for a processing system to make immediate commitments to lexical hypotheses - because of the time locked nature of speech - there are also some advantages for *delayed commitment*. Listening conditions are often far from optimal, so it might be more efficient to delay final commitment to a single lexical hypothesis instead of making many erroneous commitments which must be revised. There are a number of studies that suggest such a delay of lexical commitment to a single hypothesis.

For instance, using the gating paradigm, Grosjean (1980) found that low frequency monosyllabic words show relatively low accuracy identifications, sometimes reaching only 50% accuracy. Also, words presented in sentence context were not correctly identified until midway into the following words. Ratings of confidence did not reach 100% until approximately 350 ms after acoustic offset (Grosjean, 1985).

Zwitserlood and Schriefers (1995) investigated the contribution of processing time to lexical activation. In continuous speech, the passage of time typically results in more acoustic-phonetic information relevant to a word. This natural confound of additional time and information was separated in a cross modal priming task. A visual target was presented at the offset of an intact word or of the initial syllable. In another condition, only the initial syllable was presented followed by a delayed visual target where the delay was equal in duration to the missing word portion. They found that priming effects in the delayed, incomplete condition were comparable to those in the intact word condition (see table 5.5).

Therefore, the impact of a stretch of acoustic-phonetic information is not static and does not have (only) immediate impact on the activation level of a lexical entry. Furthermore, the evidence regarding multiple lexical activation and ambiguity preservation is consistent with this notion of delayed commitment. The results also indicate that the time-scale of conscious speech is not equal to the time scale of bottom-up processing, which is not that surprising, since activation takes time to build up, and perceptual awareness occurs after a coherent representation of the perceptual scene has been formed.

Illustrative examples of backwards effects come from *phonemic restoration* experiments (see also section 5.6.5). When a phoneme, such as /s/ in *legislature* is excised from a word and replaced by silence (*legi_lature*), subjects readily localize the silent gap. But if the silence is replaced with broadband noise, such as a cough, subjects not only fail to localize the missing phoneme, they report hearing all phonemes as present. Moreover, the context of the word and carrier sentence determines the identity of the restored phoneme. If the /s/ in "*jump on the sandwagon*" is spliced out and replaced by noise, subjects will report hearing *bandwagon*, despite the absence of the usual acoustic cues for the voiced stop consonant /b/.

Warren and Sherman (1974) found that the resolving context may be delayed for two or three, or even more words following the ambiguous word fragment. In the phrase "*It was found that the *eel is on the ----*", where the resolving context is given by the last word (*axle, shoe, orange* or *table*), listeners experience the appropriate phonemic restoration (*wheel, heel, peel*, or *meal*), apparently by "storing the incomplete information until the necessary context is supplied so that the required phoneme can be synthesized" (Warren, 1970). Thus, despite the fact that we do not perceive *orange* as occurring before *peel*, we appear to delay the formation of the *peel* percept until after the word *orange* arrives. In this example, the later occurring top-down effect of meaning influences the phonemic structure, which is consciously per-

ceived as coming earlier in time. Whether or not the phonemic restoration effects results from the kind of top-down influences that are implemented in Trace will be discussed in section 5.6.5. Nevertheless, these data clearly illustrate that the brain mechanisms that generate speech percepts can integrate contextual information across a relatively broad temporal window and still maintain a natural ordering of the linguistically significant acoustic signals that reach our ears.

These findings are consistent with those of Connine, Blasko and Hall (1991) who used ambiguous strings, compatible with two lexical items, where the initial phoneme was ambiguous and had to be labeled. Again, semantic biasing information occurred either three or six syllables downstream from the ambiguity. However, the results indicated that labeling responses were influenced by this semantic biasing information when a three syllable temporal window was used, but not with a six syllable delay. Thus, subsequent context can influence labeling responses, but there are *temporal constraints on lexical commitments*.

5.5 Lexical competition

An issue that is related to the finding that the activation of a target lexical item occurs in the context of simultaneously activated lexical candidates, concerns the consequences of this multiple activation for recognition. Specifically, whether recognition of a word is impaired when it overlaps with other words. This aspect of processing has been referred to as *lexical competition* and pertains to the idea that the activation of any given lexical candidate is affected by the presence or absence of other activated candidates that match with the current input, and which have varying degrees of frequency and phonological overlap (such words are called *phonological neighbors*).

Computational models of speech recognition all incorporate the idea of lexical competition, but differ in the way they model these competition effects. In localist connectionist models such as Trace and the selection stage in Shortlist, competition is modeled by inhibitory connections between jointly activated units in an IAC network, and the degree of inhibition is proportional to the level of activation of the lexical candidate and the degree of overlap between different candidates.

However, though behavioral data may indicate competition, it does not need to be simulated using *direct* inhibitory connections. For instance, recurrent neural networks (trained by back-propagation) produce output functions that are dependent on the number of simultaneously-activated candidates by representing the probability of all outputs given the current input. In the distributed Cohort model, lexical competition is also based on relative measurements of activation. Here, a passive, activation-based competition is proposed where recognition occurs as soon as the activation of a candidate is sufficiently greater than that of its competitors. Accordingly, the presence of a competitor has no direct influence on a target word, but its effect is only reflected in a postlexical decision stage where their relative merits are being compared. This is also consistent with the decision stage in the FLMP (section 5.2.6), where a decision is based on a relative weighting of possible candidates, which leads to a probability measure.

As the different models predict similar (behavioral) outcomes, it is difficult to distinguish between them. Nevertheless, there is one aspect in which they could be distinguished, namely the predicted patterns of activation. Passive competition models such as Cohort predict that recognition of a lexical item may be delayed in the presence of a competitor, because the decision mechanism will have to wait longer for disambiguating information to arrive, which will push the candidates' activation levels apart. But, the pattern of activation remains the same to when a competitor is absent. According to Marslen-Wilson (1993) competition has no consequences for the relative levels of activation of the competing items. Activation is

determined only by the degree of bottom-up match or mismatch. In contrast, due to mutual inhibition, an inhibitory mechanism predicts that the activation of a target word will be reduced in the presence of a competitor.

5.5.1 Ambiguity and the decay of activation

Marslen-Wilson (1990) used a gating and a cross modal priming paradigm and found that no competitor effects were found for words where a response was made on a lexical decision or auditory repetition (shadowing) task *after* the entire word. This suggests that activation of competitors decays rapidly once the lexical identity of the input is unambiguous, and that competitor effects are relatively subtle and short-lived.

However, when ambiguity remains unresolved, competitor effects do not dissipate. For instance, Marslen-Wilson et al. (1996) used cross modal priming and manipulated VOT to create ambiguous initial phonemes in words which yielded either a single competitor, such as *blank/plank* followed by *wood*, or no competitor, such as *dask/task* followed by *job*. For the *un*ambiguous source words (*plank* and *task*) virtual identical priming effects were found when followed by their semantic associate (*wood* and *job*). The priming effect of the ambiguous no-competitor prime (*d/task*) was similar to its source word. However, the ambiguous word competitor (*b/plank*) *failed* to show a priming effect (see table 5.6). It has been argued that such results reflect the signature of lexical competition and possibly, mutual inhibition. In the case of the ambiguous word competitor, the competing word forms inhibit one another comparably and form a type of "lexical deadlock", precluding sufficient activation to reach a level sufficient for recognition.

5.5.2 Inhibitory form-based priming

Results from another line of research have been interpreted as providing more direct evidence for competition between activated word candidates. These results are obtained within a *form-based priming* paradigm, where it is shown that pre-activation of a close competitor influences identification of a subsequently presented target word. Form-based priming refers to priming experiments in which the relation between primes and targets is defined in *physical* terms rather than abstract, knowledge-based terms (as is the case with *semantic priming*). In general, listeners find it harder to recognize a target word when it is preceded by a prime word sharing phonetic or phonological material. As the results of form-based priming experiments have been taken as more direct evidence for inhibitory competitive processes, some of these results will be presented.

Earlier research on auditory form-based priming seemed to reflect *facilitatory* priming effects (reflected in faster latency data (RTs), and increased accuracy (decrease in %error)). For instance, Slowiaczek and Pisoni (1987) found facilitatory priming effects in a perceptual identification task for targets presented in noise, when these were preceded by a phonologically overlapping prime (i.e., sharing at least one phoneme), presented in the clear (with an ISI[75] of 50 ms). This facilitation effect increased when the noise level was increased (from an SNR of 10 dB to –10 dB).

[75] ISI: *Inter Stimulus Interval*, the time between the offset of the presentation of the prime and the onset of the presentation of the target.

Experimental condition	Stimuli		Results*
	Auditory prime	Visual target	
Unambiguous source words			
Single competitor	*plank* (competitor: *blank*)	*wood*	Facilitatory priming
No competitor	*task*	*job*	Facilitatory priming
Ambiguous words			
Single competitor	*?lank* *?lank*	*wood*	No priming effect
No competitor	*?ask* *?ask*	*job*	Priming effect similar to unambiguous source word
Conclusion: The absence of a priming effect in the ambiguous conditions where two lexical interpretations are possible, reflects processes related to lexical competition. This might be due to mutual inhibition.			

Table 5.6 Summary of the results regarding competitive effects as reflected in the amount of priming in a cross-modal semantic priming task (Marslen-Wilson et al., 1996). The "?" reflects the ambiguity of the initial phoneme resulting from manipulation of VOT. (*The description of the results are with reference to the no-prime conditions).

Such a pattern of results is consistent with predictions of the Cohort model concerning the residual activation of various members of a prime's cohort shortly after their rejection by the word recognition system.

Using an auditory lexical decision task, Slowiaczek et al. (1986) found only facilitatory *repetition* priming effects for both words and nonwords, but not for phonologically similar primes (sharing one, two or three phonemes). This absence of facilitatory phonological priming (again with the exception of repetition priming) in lexical decision tasks was replicated in an investigation performed by Radeau et al. (1989). They used brief SOAs[76] and nondegraded stimuli, and found significant effects of *inhibitory* priming in several conditions, which increased as the phonological overlap between targets and primes increased. However, such inhibitory effects were not reflected in the results of auditory shadowing tasks. The dissociation between the effects of phonological priming might therefore reflect inherent task and/or stimulus differences: e.g., respective sizes of the response set (the entire lexicon vs. Yes/No), the use of degraded vs. nondegraded stimuli, the differences in time pressure (i.e., the absence of time constraints in the perceptual identification task). These results are summarized in table 5.7.

Another investigation, performed by Goldinger et al. (1989), was motivated by a prediction derived from the *Neighborhood Activation Model* (NAM) of word recognition which is based on the concept of activation and competition of phonetically similar words. The NAM model predicts that form-based priming should inhibit target recognition. They there-fore conducted experiments that studied the effects of priming on perceptual identification with prime-target pairs that were phonetically confusable *when presented in noise*, but shared no common phonemes. Phonetically related and unrelated prime-target pairs were compared (for examples, see table 5.8). The results indicated reliable inhibitory priming effects. Targets were identified less accurately or more slowly when they followed phonetically similar primes. However, this effect was only robust in certain experimental conditions, i.e., with a brief ISI (50 ms.) between primes and targets (not with 500 ms. ISI), and only for targets preceded by low-frequency prime words (*not* with high-frequency words).

[76] SOA: *Stimulus Onset Asynchrony*, the time between the onset of the presentation of the prime and the onset of the presentation of the target.

Experimental condition	Auditory stimuli: Prime	Auditory stimuli: Target	Results*
Perceptual identification task			
Phonological priming - ISI: 50 ms	*bull* (in clear)	*beer* (in noise)	Facilitatory priming effect, increases when noise level increases (10 dB, 5 dB, 0 dB, -5 dB, -10 dB)
Conclusion: Priming with a phonologically overlapping word, leads to residual activation of activated cohort members shortly after rejection from Cohort? (Slowiaczek and Pisoni, 1987).			
Auditory lexical decision task			
Repetition priming	*bull* (in clear)	*bull* (in noise)	Facilitatory priming effect
Phonological priming (sharing 1-3 phonemes) - ISI: 50 or 500 ms	*bull* (in clear)	*beer* (in noise)	Auditory lexical decisions not facilitated
Conclusion: Only advantage for repetition priming condition (Slowiaczek et al., 1986). Replicated by Radeau et al. (1989).			
Auditory lexical decision task			
Phonological priming (sharing 1-3 phonemes) - short SOAs	*bull* (in clear)	*beer* (in clear)	Auditory lexical decisions not facilitated. Inhibitory effects (slower RTs, increase in error), increases when phonological overlap increases
Auditory shadowing			
Phonological priming (sharing 1-3 phonemes) - short SOAs	*bull* (in clear)	*beer* (in clear)	Null result. Inhibitory effects in one experiment.
Conclusion: No replication of facilitatory priming observed by Slowiaczek et al. (1987), with the exception of repetition priming. (Radeau et al., 1989) Due to task differences and/or stimuli differences? Guessing strategy used in perceptual identification task?			

Table 5.7 Summary of the results found within the form-based priming paradigm. (*The description of the results are with reference to the no-prime conditions). For the results of Goldinger et al. (1989, 1992), see the text.

Also, *neighborhood density effects* were found. Target words with many, highly frequent neighbors are more difficult to identify when presented in noise than words with few, low-frequency neighbors (see also, Luce et al., 1990). Goldinger et al. (1989) interpreted their results as being consistent with accounts of word recognition based on activation and competition among words in memory (for example, NAM, Trace and Shortlist). In conditions that allowed enough time for activation to dissipate (longer ISI, or rapidly recognized high-frequency primes), the competition was eliminated before target presentation, and target recognition was independent of the priming manipulation.

From the described results, a distinction can be made between *phonetic* and *phonological* priming. Phonological priming of spoken words refers to improved recognition of targets preceded by primes that share at least one of their constituent phonemes (e.g. *bull-beer*). Phonetic priming refers to reduced recognition of targets preceded by primes that share no phonemes with targets but are phonetically similar[77] to targets (e.g., *bull-veer*).

[77] Such phonetic similarity is experimentally determined, and depends on the presence of noise. Without noise, phonetically similar pairs are not confusable. Phonetically similar words are typically words where one or more phonemes differ in one acoustic-phonetic feature.

	Target	Phonological prime	Phonetic prime	Unrelated prime
Perceptual identification				
All real words	*run*	*ram*	*lamb*	*bed*
	bone	*bang*	*dung*	*lease*
	mush	*mass*	*notch*	*surge*
	nerve	*gnaw*	*mug*	*whose*
	gear	*ghoul*	*yell*	*man*
	jar	*jail*	*gall*	*phone*
Lexical decision				
words	*run*	*ram*	*lamb*	*bed*
	bone	*bang*	*dung*	*lease*
	mush	*mass*	*notch*	*surge*
nonwords	*/noik/*	*gnaw*	*mug*	*whose*
	/gour/	*ghoul*	*yell*	*man*
	/jær/	*jail*	*gall*	*phone*

Table 5.8 Examples of the kind of stimuli used in the perceptual identification tasks and lexical decision tasks. (From Goldinger et al., 1992).

Goldinger et al. (1992) studied the nature of these differences by juxtaposing phonetic and phonological priming tasks using a common pool of stimuli across a variety of experimental contexts (see table 5.8). Within a perceptual identification task, they found opposite priming effects for the phonetic and the phonological condition. Phonetically unrelated primes were more accurately identified, phonetically related primes led to inhibited target identification, and phonological primes led to an overall better performance in that the targets preceded by related primes were more accurately identified.

However, a bias effect seemed to be involved. When the percentage of related trials was reduced, an error analysis indicated a significant reduce in this bias in the phonological condition, whereas the phonetic condition was unaffected by this manipulation. This implies that the facilitatory priming effects reported by Slowiaczek et al. (1987) was due to more than simple residual activation among words in memory.

The results on an auditory lexical decision task (where stimulus degradation was manipulated) confirmed this interpretation. The disparate effects of phonetic and phonological priming were only observed in noise. Additionally, on the unrelated priming trials, subjects were reliably slower and less accurate in the phonological condition when noise was present, but there was no effect of priming in the No-Noise condition. This means that the facilitatory effect entails a "cost" for the unrelated priming trials, which implies the use of a bias.

The time course of the respective priming effects was investigated by increasing the ISI from 50 to 500 and 1500 ms. As expected, reliable facilitatory effects of phonological priming were still observed (reflecting a bias that was maintained throughout the experiment), whereas no evidence for phonetic priming was found in both ISI conditions. Furthermore, when the percentage of related trials was again reduced, the influence of bias was eliminated at both ISIs (50 ms and 500 ms). In this case, a general inhibitory effect for *both* types of priming were found! Furthermore, the responses in the phonetic priming conditions were more accurate, and facilitatory effects were found for nonwords (that are not subject to competition) in both priming conditions with an ISI of 50 ms. These results suggest that the underlying mechanisms of the earlier results on inhibitory phonetic and facilitatory phonological priming are qualitatively different. They display different time courses and also different interactions with changes in proportion of related priming trials.

Goldinger et al. (1992) therefore conclude that form-based priming is essentially inhibitory in nature. Inhibition[78] is caused by bottom-up, transient competition among words in memory. Therefore, form-based priming appears to reflect basic properties of activation and competition in spoken word recognition.

On the other hand, facilitation is caused by a perceptual or a response bias, and can override the competitive effects. A *response bias* would imply a conscious and intentional response selection strategy after completion of perceptual processing. It is more likely that the locus of the bias is *perceptual*. Priming then leads to changes in resting activation levels or recognition thresholds, or criterion shifts in a perceptual decision system. Such a perceptual "assumption" (or expectancy) favors the interpretation of degraded stimuli via induced knowledge of the primes such that the bias applies only correctly on related priming trials. Such a perceptual bias is developed via implicit learning during the course of an experiment, and exerts its influence automatically. Evidence regarding the ability of subjects to implicitly learn such within-experiment regularities has been repeatedly found within many experimental tasks (e.g., Cleeremans, 2001).

5.5.2.1 *Modality differences in priming and neighborhood effects*

The previous results are all consistent with the idea of multiple lexical activation and the idea that recognition of any given word is influenced by the presence of similar words. Especially interesting are the *differences* found between the visual and auditory modality. For instance, there seems to be an asymmetry in neighborhood effects. Visual word recognition is facilitated by the presence of many neighbors. Low-frequency words from dense neighborhoods are recognized more quickly than low-frequency words from sparse neighborhoods. For high-frequency words, recognition is unaffected, presumably because their recognition is already advantaged relative to other words in the lexicon. In auditory word recognition, the opposite effect is found. High neighborhood densities inhibit word recognition, and this effect is greater for low-frequency than for high-frequency words.

There is also an asymmetry regarding the results on form-based priming. In vision, low-frequency unmasked primes interfere with the recognition of high-frequency targets, and high-frequency masked primes interfere with the recognition of low-frequency targets. So, degraded high-frequency primes inhibit target identification. In audition, low-frequency primes inhibit recognition of degraded targets. As for low-frequency words more evidence and/or additional time is required for recognition, this implies that shortly after word presentation the ambiguity is still unresolved, and multiple candidates are still active. Therefore, they can exert their influence on subsequent perceptual information processing, which may lead to interference when the new input shares part of the features with the previous input.

It is argued by Goldinger et al. (1992), that these disparate effects between the two modalities are not easily explained by assuming a unitary set of processes operating on all lexical stimuli (where speech is converted to a string of phonemes, which is then treated as a string of letters). The different effects of similarity neighborhoods may reflect different optimization strategies: The facilitatory effects of dense orthographic neighborhoods may be a by-product of a system optimized to process spatial, relatively invariant stimuli. The inhibitory effect of dense phonetic neighborhoods may be a by-product of a system optimized to process temporal, variable stimuli.

[78] If the lexical activation level directly reflects the probability of a lexical item, this inhibition effect, as reflected in longer RTs, is also consistent with a slower growth of activation (due to the presence of noise leading to more ambiguity, i.e., more lexical candidates). This slower build-up of lexical activation does not need to be the result from inhibition between lexical candidates.

5.5.3 Embedded words

Another body of research investigating the consequences of lexical competition assesses the activation of embedded words. In section 5.4.3 it was already shown that multiple word forms are activated when they are aligned or misaligned from onset (such as *bone* in *trombone*). Since these words represent different interpretations regarding a certain stretch of acoustic-phonetic information, the question is whether these embedded words enter into competitive relations.

Cluff and Luce (1990) showed that recognition of bisyllabic items comprised of separate words is affected by the neighborhood density and frequency of its constituent words. Similarly, Vroomen and de Gelder (1995) used a cross modal identity priming task (in Dutch) to asses the activation of an initial syllable word which overlapped with the onset of a second syllable. A systematic reduction in priming was found the more competitors that began with the onset of the second syllable. For example, listeners were faster in visual lexical decision to *melk* (milk) after hearing *melkem* (no second syllable competitors beginning from the /k/) than after hearing *melkeum* (few second syllable competitors). These responses in turn were faster than those made after hearing *melkaam* (many second syllable competitors) though the latter were still faster than those made after hearing a control word *lastem*. Thus, lexical candidates activated by input arriving later in time than the target word influenced recognition of the target in that the greater number of potential candidates beginning with the second syllable interferes with recognition of the initial word. However, performance was facilitated when the initial word did not overlap with the onset of the second syllable and when the second syllable had many competitors. In this case, no part of the initial word competes with alternative lexical hypotheses starting at the second syllable which in addition was part of a dense neighborhood thereby providing strong evidence for word segmentation.

Similar effects were found by Norris, McQueen and Cutler (1995) in a *word spotting* task where subjects had to detect words in nonsense strings. In another research performed by McQueen, Norris and Culter (1996), again using a word spotting task, additional evidence for competition was obtained. Words embedded in bisyllabic nonsense strings had to be detected where some of these strings were themselves the onsets of longer words. For example, subjects had to spot the word *mess* in /dəmes/, the onset of *domestic*, or in the nonword /nəmes/. It was found that target words where harder to detect in word onsets such as /dəmes/, even when subjects were told that the target would only appear at the ends of the nonsense strings. The competitor word *domestic* appeared to inhibit target recognition. Also, subjects had more difficulty detecting words misaligned from onset that consisted of real word onsets. For example, detecting *mess* in /dəmes/ was harder than detecting *sack* in /sækrif/ (from *sacrifice*). Activation of *mess* is presumably inhibited by the relatively greater activation of *domestic* which benefits from earlier activation. In the case of words aligned from onset, such as *sack* and *sacrifice*, activation is more comparable since a dominant inhibitor has yet to be established.

In keeping with the assumptions of Trace and Shortlist, these results were interpreted as providing evidence for direct inhibition, where the strength of lexical inhibition is proportional to the degree of activation afforded to the competitor and the degree of overlap between competitors. The Cohort model on the other hand, would explain these results more in terms of settling behavior in an attractor network. The activation of multiple lexical candidates leads to the system needing more time to reach a stable state. In addition, in the case of misaligned words, the earlier activation of the word *domestic* has to be overcome by the later arriving word *mess* which is not the case for *sack* and *sacrifice*. Therefore, according to the Cohort model, there is no need for direct inhibition to explain lexical competition effects.

Simulations with simple recurrent networks, where the activation values represent output probability values, also show effects of multiple lexical candidates without the need for direct inhibition, simply because the probability for selecting a certain lexical candidates is reduced in the presence of multiple candidates accounting for the input (e.g., Davis, 2000).

Nevertheless, the results found with embedded words seem to indicate that lexical candidates activated by input later in time than the target word influence recognition of the target, and that the amount of influence is determined by the characteristics of the lexical candidates (e.g., neighborhood density and/or word frequency). Apparently, contrary to the basic immediacy assumption of Cohort, processing continues after a uniqueness point such that commitment to a single lexical hypothesis is frequently delayed. This is only a real problem for Cohort if it is assumed that only the cohort representing a single (set of) lexical item(s) in the speech input can be active at the time that the input is presented.

The need for delayed commitment is strengthened by the statistics on the pattern of occurrence of words embedded in polysyllabic words of the English vocabulary (McQueen et al., 1995). A large majority of polysyllables have shorter words embedded within them (84 % when only syllable overlap is taken into account). These embeddings are most common at the onsets of the longer words. Phonological and syntactic constraints could rule out some embedded words, but according to McQueen et al., this does not remove the problem. Therefore, they argue that lexical competition provides a necessary additional means of dealing with such lexical embeddings. However, alternative statistics were a correction is made for morphologically embedded words (through affixing, such as *dark* and *darkness*, or through compounding such as in *darkroom*), reduces the proportion of embedded words in polysyllabic words to only 27% (Davis, 2000).[79]

As will be seen in the next section, alternative explanations exist to account for lexical segmentation of (onset-) embedded words without incorporating lexical inhibition. Further, concerning delayed commitment, all that is actually needed is the availability of an activated lexical item after the acoustic offset. In essence, this is similar to the availability of the activation pattern generated by previously presented acoustic input representing smaller segments such as phonemes or syllables.

Since activation gradually builds up, and dependent on the presence of mismatching information, decays slowly, this actually reflects a natural process: the activation has not dissipated immediately after the offset of the acoustic input that has initiated the neural activity pattern, but continues to build up gradually over time.

5.5.4 Word segmentation

A related argument for incorporating explicit lexical competition procedures, is that it could help in the process of word segmentation. In Trace and Shortlist, the process of spoken word recognition consists of a waxing and waning of lexical hypotheses, where the candidates that most likely describe the full body of incoming data emerge from the process of lexical competition. According to Trace, word recognition could occur in such a way even when cues to word boundaries are missing.

[79]In the earlier counts, a full-listing approach was taken in which it was assumed that the lexical representation of a word *dark* in a transparently derived word such as *darkness* is an entirely separate representation. However, experimental evidence seems to indicate that the lexical representation of the morphologically complex word is formed out of a representation of the stem combined with a representation of an affix (Marslen-Wilson, Ford, Older, and Zhou, 1996; Marslen-Wilson, Tyler, Waksler, and Older, 1994). Consequently, a lexical system that accessed the embedded word *dark* while processing a related word such as *darkness* would not be required to back-track or revise its hypothesis. In a morphologically-decomposed lexicon the presence of these onset-embeddings would help, not hinder, the recognition system. It is therefore worthwhile to revise these estimates of the proportion of words containing a lexical embedding in the light of a more realistic, morphologically-decomposed view of the mental lexicon.

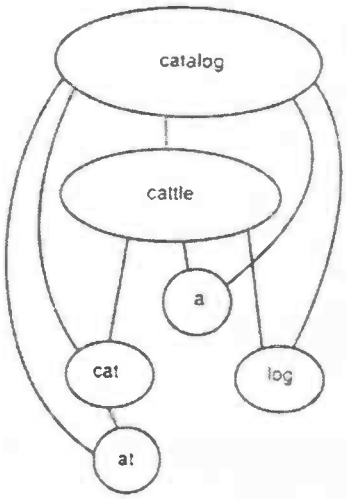


Figure 5.7 The pattern of inhibitory connections between candidates produced by presentation of *catalog*. The figure shows only the subset of candidates completely matching the input. The full candidate set would also include words such as *battle*, *catalyst*, etc. (From Norris, 1994).

There is an interesting prediction that can be derived from the way the lexical competition procedures in Shortlist and Trace are implemented (see figure 5.7). For instance, with onset-embedded words, the identification of the embedded words is delayed until longer interpretations have been ruled out, and the activation of the short word hypotheses is increased relative to the activation of the longer word in which it is embedded. This is because long words have more overlap with other words, and in addition, are more likely to encounter lexical competitors. Therefore, they are inhibited more than the shorter word that has less overlap with less competitors, which leads to an initial bias in the early build-up of lexical activation for short words. Such an initially higher activation level for short words would be needed in cases where subsequent information increases the activation of longer words, but not for the shorter words. Therefore, the bias towards short words supports the identification of embedded words since it provides them with the additional activation required for them to win out in competition with longer lexical items.

Conversely, recurrent networks predict another pattern of activation, namely that embedded words and longer competitors will be equally activated where both words match the speech stream. Multiple candidates are each activated in proportion to the conditional probability of that item being present in the current input, irrespective of length (see figure 5.8 and 5.9).

5.5.4.1 Bottom-up mismatch information

The apparent need for a short word bias in Trace can be ascribed to the fact that McClelland and Elman (1986) do not incorporate bottom-up inhibitory connections between, for instance, phoneme units and lexical units representing words that do not contain those phonemes. As Trace uses inhibitory connections at the lexical level to rule out potential candidate words, the model is only able to reduce the activation of a lexical item following mismatch where there are other more active units that can provide inhibition at the lexical level. Consequently, Trace predicts that in cases where input that mismatches with an activated candidate is presented, there should be little or no decrease in lexical activation if the input does not match an alternative lexical item.

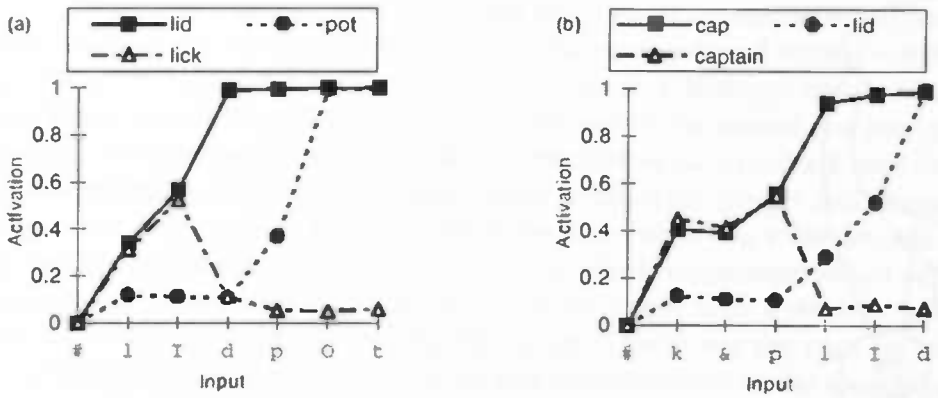


Figure 5.8 Activation of cohort competitors and onset-embedded words: (a) Cohort competitors (*lid/lick*) during the sequence "*lid pot*", (b) Onset-embedded words (*cap/captain*) during the sequence "*cap lid*". From Davis (2000).

This prediction was not borne out by experiments reported by Marslen-Wilson and Gaskell (1992, 1994), since mismatch that creates a nonword (e.g. *sausin* mismatching with *sausage*) reduces priming to an associatively related target as effectively as mismatch that produces an alternative word (*cabin* versus *cabbage*). Gaskell (1994) shows that the standard Trace model is unable to account for this data and suggests that accounts (such as Cohort and Shortlist) incorporating bottom-up inhibition (such that the mismatching final segment of *sausin* reduces the activation of *sausage*) may provide a better account of effects of mismatch on lexical activation than models that rely solely on intra-lexical competition.

Davis (2000) studied the time-course of identification of onset-embedded words and longer competitors in order to test the alternative accounts, i.e., accounts with or without direct inhibition. Using the gating paradigm and cross-modal repetition priming, Davis investigated the recognition of sentences containing a long word (*captain*) or an embedded word in a context which matches (*cap tucked*) or immediately mismatches (*cap looking*) with longer words. By using, for instance, the magnitude of priming as a measure of activation, it was shown that the perceptual system can already distinguish between monosyllabic words and the onset of a bisyllable even at the offset of the first syllable.

Davis suggests that acoustic (sub-phonemic) cues can be used by the perceptual system to allow such early disambiguation. For instance, the difference in (relative) syllable length as indicated by acoustic-phonetic analyses of syllable duration for embedded words and short words (especially syllable lengthening in monosyllabic words). Though previous results indicated identical priming effects for *cap* and *captain* after the presentation of *cap*, these studies used single word presentations. Acoustic cues to relative syllable duration - also related to speech rate - can not be estimated under these circumstances and can therefore not aid in disambiguating between the two interpretations. Presenting words in sentence context, however, does provide the means to use such cues. Therefore, differences in the acoustic form of syllables from short and long words can directly affect relative levels of lexical activation for the short and long target words.

Another suggestion as to how the speech recognition system might proceed in the process of word segmentation, is that recognition operates in a strictly left-to-right fashion. Such an account seems to require that words can reliably be identified before their offset. The analyses presented earlier suggest that many words do not diverge from other lexical candi dates until after their final phoneme (Luce, 1986). These are words that are embedded at the onset of longer lexical items (such as *cap* embedded in *captain, captive* etc.). However, such analyses are based on phonemic mismatches. In an on-line model such as Cohort, it is assumed that sub-phonemic cues are immediately used to differentiate between lexical candidates (see also the featural account of lexical access and representation described in section 5.3.3).

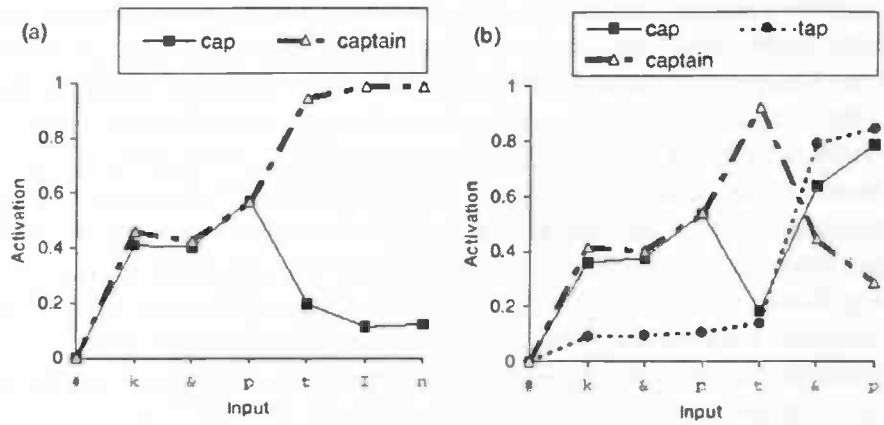


Figure 5.9 Activation of embedded words and competitors for long word sequences and lexical garden-paths from simulations with a recurrent network. (a) Bisyllables with embeddings (*captain/cap*) during "*captain*", (b) Lexical "garden paths" (*cap/captain*) during "*cap tap*". From Davis (2000).

The experimental evidence presented by Davis (2000) indicates that the perceptual system indeed uses subphonemic cues to distinguish onset-embedded words from the start of longer words in which they are embedded. To the extent that this is generally the case, processes such as delayed recognition and mechanisms of lexical competition may play a less important role in spoken word recogni-tion than would be argued in accounts based on a purely phonemic analysis of the speech stream. Indeed, it seems that the crucial property of systems like Trace and Shortlist is not the inclusion of direct intra-lexical competition, but the fact that the recognition system is not confined to using sections of the speech stream to identify only a single word at a time. For instance, systems in which the goal of the recognition process is to activate a representation of an entire sequence of words are also capable of using post-offset information to identify onset-embedded words (Davis, 2000).[80]

5.5.4.2 *Statistical accounts*

An important class of theories of segmentation proposes that the statistical or distributional properties of lexical items can be used as a cue to the location of word boundaries. Such accounts are particularly popular in the developmental literature, since they suggest ways in which infants might learn to identify words in connected speech without any obvious cues to determine where boundaries between lexical items are located. The extent to which distributional information is utilized by adults - who have lexical knowledge to bring to bear on the segmentation problem - is unclear. It has been suggested that the processes of lexical access and identification will be much more efficient if the recognition system can reliably identify word onsets prelexically (see Briscoe, 1989). The argument is that if a prelexical segmentation strategy is used then instead of initiating frequent, unsuccessful lexical access attempts the recognition system can ensure that fewer inappropriate lexical access attempts will be made. The right segmentation strategy would therefore allow more efficient recognition without compromising accuracy.

[80] An interesting assumption made by such systems is that the task involved in learning to understand spoken language involves a mapping from whole utterances of connected speech to the entire meaning of that entire sequence. Therefore, there is no need for a one-to-one correspondence between units in the speech stream and units of meaning. Lexical representation emerges as the fundamental unit of regularity in the mapping between speech and meaning, instead of forming the basis for a form-meaning mapping. A related approach is the one described by St.John and McClelland (1990) regarding a model of sentence processing, where the goal of their model is to activate a 'sentence gestalt'; a representation capturing the thematic relationships between constituents in a sentence.

One cue that has been proposed as a lexical segmentation strategy is metrical structure (such as the rhythmic alternations between weak and strong syllables in English). This has been incorporated in Shortlist by means of a Metrical Stress Strategy (MSS) (Norris, McQueen and Cutler, 1995). Thus, in addition to the lexical competition process, a second process, in which sensitivity to metrical structure is incorporated, aids in word segmentation. At the selection stage, possible new lexical candidates can get additional activation, i.e., segmentation of the speech input occurs at the onset of strong syllables and in addition, a lexical access attempt is initiated from the points at which segmentation has occurred. As the strong syllable occurs after the offset of the previous word, such a strategy introduces discontinuities in processing and requires backtracking. This is not the case for models in which it is assumed that listeners construct an on-line interpretation of the speech signal as it unfolds by using prelexical cues (e.g., Cohort).

The primary value with a stress-based strategy lies in the fact that, in typical English speech, more than 90% of content words do begin with a strong syllable, and approximately 75% of all strong syllables are indeed the initial syllables of content words (Cutler and Cartler, 1987). However, this strategy would only work for content words, as the reverse (weak-initial) pattern is found for *closed-class words*: 69% of weak syllables are at the onsets of closed-class words, with fewer than 5% being the initial syllables of *open-class words*.[81]

Nevertheless, it has been shown that listeners (even infants at the age of 9 months) are sensitive to rhythmic patterns in speech, and experimental evidence indicates that listeners make use of the (implicitly known) regularities of their native language to perform tasks such as word-spotting. For instance, it has been shown that English listeners are faster to detect monosyllables followed by a strong syllable than by an unstressed (weak) syllable (Cutler & Norris, 1988; Norris, McQueen, & Cutler, 1995). Another problem with MSS is that, in order to place word boundaries, it is required that listeners are able to detect syllable boundaries prior to lexical access. While sonority hierarchies do provide a preliminary grouping of segments into syllabic units, phenomena such as re-syllabification, whereby a sequence such as *band ate* will be syllabified as *ban date*, mean that word boundaries will not necessarily fall at syllable boundaries.

Another approach shown to be effective for the identification of syllable and word boundaries uses distributional or phonotactic regularities in segment sequences. The use of distributional regularity (DR) as a cue to lexical segmentation follows the assumption that chunking the speech stream into frequently occurring sequences will extract linguistically coherent units. For instance, it has been shown that algorithms, which operate by minimizing the description length of a corpus of utterances are able to extract lexical items contained in it. Such an algorithm compares different sets of lexical items that can be used to transcribe the utterances in the corpus, and chooses the set that uses the minimum number of lexical items, while minimizing the total length of these lexical items and maximizing the product of the frequency of occurrence of each lexical item. The lexicon discovered by this distributional regularity (DR) algorithm for a phonologically transcribed corpus of child-directed speech corresponds fairly closely to the words contained in the orthographic transcription of this corpus. Performance was further improved by providing phonotactic constraints on the system's segmentations (Brent & Cartwright, 1996).

Similar systems have also been developed using on-line learning in a neural network. For example, Elman (1990) used an SRN to predict the next input segment in a small artificial corpus. Elman reports that output error drops as the network is presented with more of a word in the training set and rises sharply at the offset of each word. Information from this "saw-tooth" error could therefore be used to determine which sequences of input segments

[81] *Closed-classed words* are function words such as pronouns, conjunctions, and determiners. Newly coined or borrowed words can not be added to these categories. The *class of open words* (content words) contains the major lexical categories: nouns, verbs, adjectives, and adverbs.

constitutes a word in the language that the network is exposed to. The network's output error will be high where multiple phonemes can follow the current input, i.e., at a word boundary. Thus, the system can be used to determine which sections of the speech input cohere as linguistically salient units.

Cairns et al. (1997) extended this recurrent network account to a larger corpus transcribed into a distributed phonological representation. They found that when using the increased error as a cue to possible word boundaries the network successfully identified 21% word boundaries in a test set. The network seemed to make no distinction between syllable and word boundaries. This indicates that the network actually extracted phonotactic constraints that allow the detection of boundaries between well-formed syllables, not boundaries between words. The success of this approach in detecting word boundaries seems to reflect the fact that many words in English (and high frequency words in particular) tend to be monosyllabic. Therefore, this result does not have to be viewed as entirely negative. Since the majority of early acquired words in English are mono-syllabic (Aslin, Woodward, LaMendola, & Bever, 1996), such an approach provides an interesting account of how the infant comes to "bootstrap" segmentation prior to lexical acquisition.

An account that is closely related to these distributional regularity theories involves the use of phonotactic information as a cue to lexical segmentation. Phonotactic accounts are based on the same assumption that chunking the speech stream into frequently occurring sequences extracts linguistically coherent units. However, the procedures used in phonotactic accounts of segmentation take the opposite approach as they are looking for word boundaries rather than words. Phonotactic accounts assume that *in*frequently occurring sequences are likely to contain boundaries between distinct linguistic units. Furthermore, when phonotactic constraints are explicitly incorporated, such as the requirement that illegal phoneme sequences (that never occur within words) must contain a word boundary or the requirement for the presence of a vowel in each word, the performance obtained with neural network simulations was more successful. Off course, infants learning their language are not presented with such explicit phonotactic constraints in order to segment continuous speech, but developmental research on language acquisition has indicated that 8-month-old infants are already very sensitive to the phonotactic regularities of their native language.

To conclude, a variety of different forms of statistical regularities in the speech stream can be used as a cue in learning to segment connected speech. Each of the cues proposed has been shown to be effective in computational simulations, and also has evidence to support its use in infants' segmentation of connected speech prior to the acquisition of lexical items. Recurrent network simulations reported by Christiansen, Allen and Seidenberg (1998) investigated the strength of combining different combinations of these cues. Not surprisingly, they showed that *best performance is obtained by combining multiple sources of constraint* in a single network.

Therefore, though no source of information is in itself expected to be sufficiently reliable, a combination of the following sources of information should provide a sufficient means to accomplish word segmentation:

- statistical regularities derived from previous language exposure, which become implicitly encoded as a result of language exposure
- sensitivity to relevant (prelexical) acoustic cues to which listeners have learned to attend, and that allow for an efficient on-line discrimination between multiple candidates where possible
- sensitivity to prosodic information, such as rhythmic and intonation patterns, also resulting from learning about (the native) language characteristics
- lexical "knowledge" that has already been acquired during the ongoing course of language acquisition; once some coherent (lexical) "units" have been formed, they can be used to help further segmentation of the speech input, and to learn new words

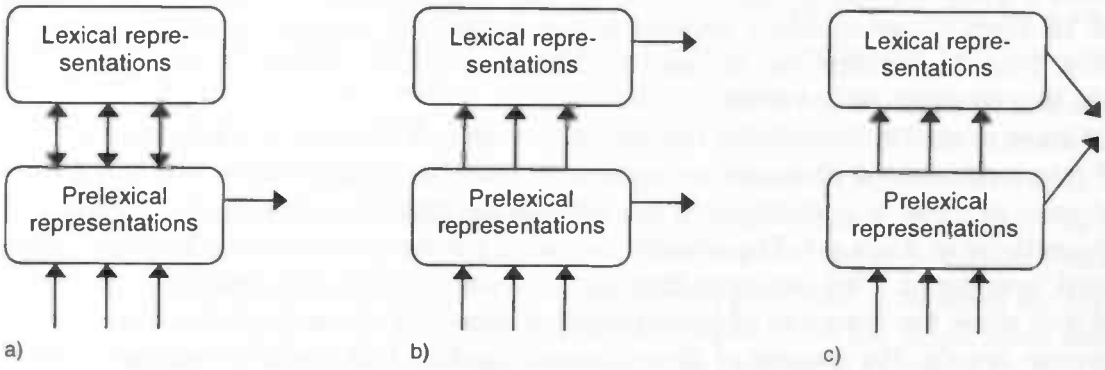


Figure 5.10 Schematic representation of how different models account for the results on phonemic decision tasks. (a) In Trace, bottom-up feature information in combination with top-down lexical information activate phonemic units. These are explicitly accessible for the purpose of making phonemic decisions. (b) In Race, prelexical and lexical representations can both provide information on phoneme identity. Phonemic decisions are based on reading out one of the outlets. Which outlet is used, is determined by a competitive race, where the earliest available information determines the representation on which phonemic decisions are based. (c) In Merge, information from prelexical and lexical representations is merged to reach a decision on phoneme identity.

- syntactic structure
- semantic constraints, related to world-knowledge, and therefore dependent on the interaction and previous experience with the environment

Notice that models such as Trace, Shortlist and Cohort describe the processing of an *adult* human speech recognition system, which has already learned to make efficient use of these language characteristics, for a large part by implicitly encoding such information. It seems likely that the need to incorporate direct inhibitory processes between lexical units in Trace and Shortlist is forced by the choice for localist representations within an IAC architecture. To simulate the competitive *behavior*, it is necessarily to "compensate" for the fact that the build-up of lexical activation is not influenced by prelexical knowledge, and that cues to segmentation such as the ones described above are not included. Also, the absence of bottom-up mismatch information in Trace overestimates the amount of lexical activation.

5.6 The direction of information flow

The final issue that will be addressed in this chapter is concerned with the direction in which information flows from different levels of representation. In models of spoken word recognition that are consistent with a parallel activation account, the interaction "versus" autonomy distinction can be interpreted as constraints on the direction of connectivity and types of connections that can be made to and from lexical units (see also section 5.2.1). Therefore, lexical influences on prelexical processes and representations, or more specifically, the presence of feedback connections from word representations to earlier representations, are of special importance. This question is typically being addressed within the context of experimental tasks in which listeners are required to make phonemic decision (e.g., phoneme monitoring, phonetic categorization, or phoneme restoration). In figure 5.10 the most important differences that exist between accounts explaining the results on phonemic decision tasks are schematized. Some of the most illustrative experimental findings will be discussed in the following sections.

5.6.1 The Ganong effect

Linguistic context has long been known to aid and bias the identification of speech. A typical example of how lexical knowledge influences phoneme perception was illustrated by Ganong (1980), who investigated the interaction of auditory information and lexical knowledge in speech perception. Using a phonetic categorization task, it was found that phoneme identification was influenced by the lexical status in which the target phoneme occurred. Ganong generated ambiguous phonemes on a continuum between a word and a nonword (e.g. *type-dype*), and found that subjects are biased towards classifying phonemes in the ambiguous region (i.e., the phoneme boundary), so as to be consistent with a word (*type*) rather than a nonword (*dype*), which is reflected in more "*t*" responses in the *dype-type* vs. *dice-tice* continuum. This basic lexical effect has been known as the *Ganong effect*. The fact that this lexical effect occurred only at the phoneme boundary was taken as evidence that lexical knowledge can influence the interpretation of acoustic information before phonetic categorization takes place. According to Ganong (1980), this would be consistent with criterion-shift models with a precategorical bias towards hearing words, whereas in a categorical perception model a lexical bias would act postcategorically by means of a correction process.

Nevertheless, it is also possible to explain the results by means of a postperceptual response strategy which results in responses that are more consistent with lexical items (Norris, 1982). According to Norris, the output of lower-level process may be quite detailed. If the early stages of processing are unable to resolve ambiguous input, they simply need to provide a rich enough output that later stages can make the correct interpretation (based on information that is available to them, but not to the lower stages).

Pitt (1995a) also studied the joint influence of phonological information and lexical context in the phonetic categorization task, and improved on earlier studies by collecting enough observations to analyse individual's data quantitatively. He studied the lexical effect on the categorization of the phonemes /g/ and /k/ in the continua *gift-kift* and *giss-kiss*. The obtained data were transformed in *d'* values by converting each subjects' proportion of /g/ responses for each step of both continua into *z* scores and then subtracting adjacent *z* scores. When plotted, the two continua differ in the location of their peak *d'* score. Pitt interpreted this finding as providing evidence for an interactive feedback system: since the lexical information shifts the *d'* measure, it must have an effect on phoneme perception. This would be consistent with the explanation given within the interactive Trace model (McClelland and Elman, 1986). When listeners are presented with input containing ambiguous phonemes, top-down activation from the lexicon will act to bias the interpretation of the ambiguous phoneme so that it is consistent with a word rather than with a nonword, exactly as observed by Ganong (1980).

However, Massaro and Oden (1995) argued that care should be taken when analyzing the performance of models against *transformed* results instead of *observed* behavior. They showed that when their FLMP was evaluated against Pitt's data, it could account for the observed pattern of behavior. Therefore, they concluded that context-dependent perception can be well explained in terms of stimulus information and context information making independent but joint contributions to word recognition.

5.6.1.1 Architectural independence

As has been pointed out by Norris et al. (2000), a problem for the FLMP is that the assumption of architectural independence becomes impossible to sustain when we consider how to account for lexical involvement in phonemic decision-making. In any model of word recognition the degree of support for a lexical hypothesis must be some function of the degree of

support for its component segments: if there is good perceptual evidence for a /g/, for example, there is also good perceptual evidence for words containing /g/. Nevertheless, in FLMP, support for /g/ in a lexical context (e.g., the extent of support for /g/ due to the word *gift* given the string /?ɪft) does not depend on whether the string begins with an unambiguous /g/, an ambiguous phoneme, or an unambiguous /k/ (Massaro & Oden 1995). Since this evaluation of contextual support for /g/ is made through comparison of the input with the stored representation of the word *gift* , this implies that information consistent (or inconsistent) with the initial /g/ does not influence the goodness of match of the input to the word. In other words, architectural independence depends on the remarkable assumption that the support for a word has nothing to do with the perceptual evidence for that word.

Other autonomous models do not assume independence in this sense. For instance, in phonemic-decision models such as Race (Cutler and Norris, 1979) and Merge (Norris et al., 2000) the activation of a word depends on the activation of the prelexical representation of its constituents, but the activated word does not subsequently impose its activation on these constituents through feedback. Prelexical evaluation therefore proceeds independent of lexical context, and lexical context effects occur because prelexical and postlexical information both exert their influence on the perceptual outcome. In Race, this is achieved by an increasing probability that phonemic decisions are based on phonological codes, whereas in Merge the information from both sources is merged to reach a decision regarding phoneme identity. Thus, although the influence of lexical information on phonetic categorisation may be simulated through top-down interactions between lexical and prelexical information such as in Trace, the results need not imply top-down connectivity. It only implies that subjects' *responses are made from a level of representation that includes lexical influences.*

5.6.1.2 *Perceptual versus belief bias*

They further emphasize that interpretations based on SDT analyses (see appendix A) are not as clear cut as has generally been assumed. Perceptual and postperceptual processes are not that easily separated: the sensitivity measure *d'* can not be equated with perception, and the decision bias measure with postperceptual guessing. The bias measure can also reflect perceptual processing. Therefore, a distinction has to be made between two types of biases, a *belief* bias and a *decision* bias. The former refers to the perceptual interpretation of the stimulus, and the latter refers to the participant's inclination to respond.

Connine and Clifton (1987) showed that these biases are different in their influence on performance. Again, subjects had to label phonemes on a /d/-/t/ continuum which were lexically biased, and the results indicated that lexical knowledge was used since listeners labeled with more identifications that formed words. This influence of lexical status was again confined to those stimuli that were perceptually ambiguous. An RT analysis of word and nonword responses also indicated a word advantage, i.e., the pattern of RTs of word judgments were faster *only* for speech stimuli that gave a lexical context effect on response probability. According to Massaro and Oden (1995), this reflects a belief bias. However, though in the monetary payoff condition a comparable labeling shift was found when listeners received a larger cash reward for a "*d*" or "*t*" response, there was *no* RT advantage for the biased response. RTs were always faster for the bias-consistent alternative, thereby reflecting a decision bias. Thus, with different payoff contingencies, the pattern of RTs for the two tasks differed even though the response probabilities did not. It seems therefore that the results that are reflected by a criterion-shift in the identification functions obtained within the phonetic categorization paradigm, do not just reflect perceptual processing, but might also reflect postperceptual response strategies.

Connine and Clifton interpreted their results as being consistent with the idea that perceptual processes compute a representation from the speech signal in concert with feedback from a lexical representation thereby supporting the assumptions of Trace. Massaro and Oden (1995), on the other hand, consider these findings as reflecting differences between a perceptual belief bias and a postperceptual response bias, thereby emphasizing that a perceptual bias does not need to have any influence on the sensitivity of early perceptual encoding processes. Though it may lead to faster processing based on less acoustic evidence, this does not need to result from a more clearly, quickly and completely encoding of the auditory input (see also Norris, 1995; Farah, 1989; Massaro and Cohen, 1991; Massaro and Cohen, 1995).

5.6.2 Lexical inhibition of phonemes

A prediction that can be derived from the Trace model is that the top-down excitatory lexical feedback in combination with the inhibitory processes at the phonemic level leads to lexically driven inhibition of phonemes. Using a phoneme monitoring task, Frauenfelder et al. (1990, Experiment 2) tested whether this prediction was borne out, and they found no evidence for such indirect lexical inhibition. In their experiment, three kinds of target stimuli (in French) were used:

(1) targets that were presented after a word's uniqueness point: /l/ in *vocabulaire*
(2) targets that were presented in derived nonwords, so-called inhibiting nonwords (INW): /t/ in *vocabutaire*
(3) targets that were presented in control nonwords (CNW): /t/ in *socabutaire*.

Instead of having longer RTs, targets in inhibiting nonwords were detected just as fast as targets in control nonwords, though both were slower than targets presented in word context (such a word advantage is consistent with earlier findings, e.g., Foss and Blank, 1980; Rubin, Turvey and van Gelder, 1976). This finding was replicated (in English) by Wurm and Samuel (1997, Experiment 1). Nevertheless, based on this null result, it is not possible to conclude that lexical inhibition is not present.

Therefore, Wurm and Samuel (1997) examined whether attentional factors at a prelexical level could have influenced the experimental outcome. Since it has been shown that listeners can readily learn within-experiment regularities, such as the target being often located near the middle of the stimuli, leading to an (automatically initiated) attentional focus at these target positions (e.g., Pitt and Samuel, 1990), the stimulus items were more balanced to correct for such possible effects. Furthermore, the conditional target probabilities, i.e., the probability of a nonword containing the target versus the probability of a word containing the target, could also have had an influence on listener's response strategies. This was also equalized. Finally, true nonwords (TNW) were used as control nonwords (for instance, *agadmakion, adunupake, doslistia*). In the Frauenfelder et al. study minimal nonwords, such as *socabutaire* (where the fricative /s/ replaces the fricative /v/), were used. It has been shown by Connine, Blasko and Titone (1993) that such minimal nonwords are not sufficiently different to block lexical activation, and in fact lead to a comparable degree of lexical activation of CNWs compared to INWs. When the stimuli were corrected for these factors, it was found that RTs to INWs were actually *faster* than RTs to TNWs.

This is an effect that is opposite to the prediction from Trace. However, instead of concluding that this provides evidence against Trace, this only indicates that attentional factors can facilitate processing of word-like nonwords such as to cancel out possible inhibitory effects. Subsequent experiments in which a dual-task paradigm was used to control attentional influences, indicated that this could indeed explain the inconsistency between Trace's

prediction and the experimental results. Nevertheless, there is still no evidence for direct lexical inhibition effects (Norris et al., 2000).

Another issue to bear in mind is that the claim that Trace predicts inhibition from the lexicon is specific to the particular implementation of Trace rather than true of interactive models in general (Peeters, Frauenfelder & Wittenburg 1989). Trace could be modified to incorporate at the phoneme level a priority rule similar to Carpenter and Grossberg's (1987) "two-thirds rule" in the ART model. In the context of a simple interactive activation model, this would mean that top-down activation would only have an effect when at least some bottom-up activation was present. That is, feedback from lexical to phonemic nodes would be contingent on there being at least some perceptual support for the phoneme. The input *vocabutaire* would then not activate /l/ at all, and /l/ would therefore not inhibit /t/.

5.6.3 Phonotactic regularities

Another context effect is the influence of phonological constraints in speech perception. This issue was addressed by Massaro (1989a) and Massaro and Cohen (1983). The stimuli consisted of phonological segments ambiguous between /l/ and /r/ that were presented in different contexts. For example, a consonant cluster syllable (CCV) beginning with one of the following consonants /p/, /t/, /f/, /v/, or /s/, followed by the ambiguous glide consonant ranging on a equally-spaced continuum from /l/ to /r/[82], followed by the vowel /i/. Subjects had to respond whether they heard /l/ or /r/. Results indicated that they were influenced by both the glide consonant and the context in such a way as to be consistent with phonological regularities (in English). For instance, there were more /r/ responses in the context /s?i/ than in the context /t?i/, where /?/ represents the ambiguous phoneme, since /l/ is permissible in /s_i/ and /r/ is not, and /r/ is permissible in /t_i/ whereas /l/ is not. In contexts where both are permissible (/f_i/) or where neither are permissible (/v_i/), no bias appeared. Though in some contexts the phonotactic acceptability was confounded with the actual lexical status of the item, i.e., /fli/ and /fri/ (*flee* and *free*), and /tri/ (*tree*), the bias effect found in the /s_i/ condition was interpreted by Massaro and Cohen (1983) as reflecting the presence of phonotactic rules. It was further shown that Massaro's FLMP was able to reproduce the behavioral data by assuming that context and stimulus information provided independent sources of information.

However, the Trace model also predicts these effects without explicitly incorporating phonotactic rules. Since the contexts that are consistent with phonotactic rules also partially activate a number of words in the mental lexicon starting with, for instance, /sli/, these contribute to the activation of /l/. This feedback support allows /l/ to dominate /r/, just as it would if /sli/ were an actual word. Regarding the independence between stimulus and lexical context, however, the original Trace model can not simulate this, since it overestimates the effects of lexical context on stimulus perception.

When the model is modified to be stochastic rather than deterministic (i.e., variability in the input to the network or intrinsic variability in the network itself), and when a best-one-wins decision rule is used (instead of a relative goodness rule, based on Luce's choice rule (Luce, 1969)), Trace also seems to be able to simulate Massaro's results. Since the modified stochastic interactive activation and competition (SIAC) model is also able to simulate the same independency, this indicates that interactive models should not be viewed as alternatives to classical accounts, but as hypotheses about the dynamics of information processing that lead to the global asymptotic behavior that the classical models describe (McClelland, 1991).

[82] The glide was changed from /l/ to /r/ by changing its initial third formant (F3) from high to low.

Also, the critical point is the falsification of interactivity itself, i.e., bi-directional propagation of information, rather than just some specific model implementing it. But, as has been pointed out by Forster (1979), it is not possible to prove the null hypothesis that interaction does not take place. Therefore, if other dynamic autonomous (bottom-up) models exist that can also account for the experimental results, according to *Occam's Razor*, these should be preferred (Norris et al., 2000).

How then can autonomous models account for the (apparent) independent influences that stimulus and context both exert on the perceptual outcome? It has already been mentioned in section 5.5.4.2 that *humans learn and become sensitive to the statistical regularities in the input*, such as the ones that can be described by phonotactic rules. These are implicitly encoded within (and between) levels as a result of learning without the need for feedback providing this information during processing. This is exactly in line with bottom-up models such as Cohort and Shortlist, and simulations have shown that they can predict the same pattern of behavior (e.g., Norris, 1993). Additional evidence to support this line of reasoning will be presented in the following sections.

5.6.4 "Compensation for coarticulation"

A study on coarticulation effects, performed by Elman and McClelland (1988), was initially taken to provide strong evidence in favor of interactive models. In order to distinguish the predictions made by interactive and autonomous models, they examined whether putative *interlevel* phenomena can trigger the operation of *intralevel* processes at lower levels. The intralevel process involved the perceptual compensation for the coarticulatory influences of one speech sound on another. It has often been found that the context in which a phoneme occurs restructures the cues to the identity of that phoneme. Several approaches to deal with this issue could be taken, such as (1) to find relatively invariant - generally *relational* – features, or (2) to redefine the phonemic unit, so that it encompasses the context and therefore becomes more invariant, or (3) a combination of these two, i.e., to use (less sensitive) *dynamical* representation in which contextual factors are already implicitly encoded.

However, a general approach that is taken within Trace to account for the context-sensitivity of cues, is to allow context to retune the perceptual mechanism on the fly: the perceptual system uses information from the context in which an utterance occurs to *alter* connections. Trace therefore predicts that this compensation can be triggered by illusory phonemes which are perceived as a result of top-down lexical influences.

Trace's explanation of the Ganong effect given earlier was that the lexical bias alters the activation of the component phonemes. This means that for instance an ambiguous phoneme /?/ midway between /ʃ/ and /s/ will activate the /ʃ/ phoneme node in the context of *fooli?* and the /s/ node in the context of *christma?*. Based on a finding reported by Mann and Repp (1981), this in turn would lead to a shift in the phoneme boundary between /t/ and /k/ that is closer to /k/ following /ʃ/ (reflected by more /t/ responses) and that is closer to /t/ following /s/.[83]

[83] As has been mentioned in chapter 2, coarticulatory influences result from vocal tract dynamics and are a major source of variability in the speech signal. Listeners "compensate" perceptually for the coarticulatory influences of one phoneme on the production, and hence acoustic realization, of its neighbors (Repp and Liberman, 1984). An example can be seen in the effect of the phonemes /s/ and /ʃ/ on the pronunciation and perception of subsequent stop consonants (e.g. /d/, /t/, /g/, and /k/). The phoneme /ʃ/ is in *ship* is pronounced with a rounding of the lips, which causes an elongation of the vocal tract that may persist through neighboring phonemes (e.g. in *foolish dancer*, the effect of the /ʃ/ colors the sounds /d/ and /æ/). In contrast, the phoneme /s/ is produced by retracting the lips, thereby shortening the vocal tract. Thus, the /d/ in *fearless dancer* is produced with a shorter vocal tract than the /d/ in *foolish dancer* which results in a shift of the distribution of acoustic energy into a lower energy range. (Note is continued on the next page).

Therefore, an effect of "compensation for coarticulation" would be produced for the ambiguous phoneme based on top-down information that would be indistinguishable from when the listener had really heard either /s/ or /ʃ/. This is because it is believed that one of the essential functions of the speech perception system is to factor out the contextual influences and recover the underlying phonetic code (Elman and McClelland, 1988; Fowler, 1985; Liberman et al., 1967). Indeed, this is exactly what the findings reported by Elman and McClelland indicated. This has therefore been taken to support a central pre-mise of interactive models, in which basic aspects of perceptual processing are subject to influences from higher levels.

However, another interpretation would be that the results are due entirely to effects operating at a prelexical level. For instance, simulations with SRNs having to learn to make decisions about phoneme identity for phonemes in contexts, show that such a network (where no word nodes are present) is able to learn the statistical properties of the language, i.e., it learns about the context in which certain phonemes are most likely to appear in (Norris, 1993; Cairns et al.,1995). Such within-level statistical information also leads to apparent interactive effects in these simulations.

Cairns et al. (1995) also analyzed a large corpus of spoken English and showed that after /ə/, /s/ is more likely than /ʃ/, and after /ɪ/, /ʃ/ is more likely. All of Elman and McClelland's (1988) /s/-final words ended /əs/, and all of their /ʃ/-final words ended /ɪʃ/. Since their material contained sequential probability biases that could in principle be learned at the prelexical level, their results can not distinguish between interactive and autonomous models.

Therefore, Pitt and McQueen (1988) used nonword contexts ending with unambiguous or ambiguous fricatives that contained transitional biases, followed again by a word-initial /t/-/k/ continuum. Again, the probability bias was reflected in the categorization of the ambiguous fricative, accompanied with a shift in the identification function for the following /t/-/k/ continuum. This suggests that compensation for coarticulation was being triggered by the probability bias rather than to the effects of specific words. In addition, when two word contexts were used (*juice* and *bush*) where the transitional probabilities of /s/ and /ʃ/ were matched, no shift was found in the stop identification function following *jui?* and *bu?*. This suggests that the compensation for coarticulation mechanism is immune to effects of specific lexical knowledge.

However, there were lexical effects in the identification of the ambiguous fricative (more /s/ responses to *jui?* than to *bu?*), which reflects a dissocation in the word context between the lexical effect observed in fricative labeling and the absence of a lexical effect in stop labeling, which is inconsistent with Trace's predictions. Therefore, sensitivity to sequential probabilities should be modeled at the prelexical level (see also Vitevich and Luce, 1998). These results also illustrate how *low-level statistical properties of the language can give rise to effects which can easily masquerade as top-down influences.*

Furthermore, since /s/ has an alveolar place of articulation whereas /ʃ/ is a palatal sound, the front cavity is shorter in the case of /s/ and its spectrum is higher. This may pull the place of articulation of nearby velar, and perhaps, alveolar stops forward, causing them to have higher spectra than in other phonetic environments (Repp and Mann, 1980). It can be illustrated that listeners compensate for these effects by adjusting the boundary between phonetic categories which are distinguished by the frequency distribution of acoustic energy. When for instance a sequence of sounds is constructed, ranging from /t/ to /k/ or from /d/ to /g/, by progressively lowering the distribution of acoustic energy contained in the noise burst that occurs on release of the stop consonant. As a result, an ambiguous phoneme between /t/ and /k/ is interpreted more often as /t/ after /ʃ/ , but as a /k/ after /s/.

Other accounts for how some apparent top-down interaction can arise in a strictly bottom-up model (e.g., Fodor, 1983; Forster, 1981) also emphasize that frequent sequences of events result in stronger intramodule associations for these sequences leading to a facilitated "filling in" of degraded or missing input. This facilitation originates from *intramodule dynamics* instead of higher-level information. Such influences are the result of prior learning, where the system has adapted to the environment such as to adequately and efficiently represent the environmental regularities. So, *the nature of the processing interactions that develop between different representations depends on the regularities that exist between different domains, and only indirectly on the assumptions made by the modeler.*

5.6.5 "Phoneme restoration"

Another frequently replicated experimental effect that has often been taken as strong evidence for interactive models of spoken word recognition, is the phoneme restoration (PR) effect, which is a powerful auditory illusion (for a review, see Samuel, 1996). When part of an utterance is replaced by another sound (e.g., white noise or speech-correlated noise), listeners report that the utterance sounds intact – they perceptually restore the missing speech and perceive them as being "real". They have also difficulty in determining where in the speech the noise occurred. The phoneme restoration effect is not a perceptual effect that stands in itself. Lots of other perceptual phenomena are known where missing evidence seems to be "filled in" by the perceptual system. In general, this has been called the continuity illusion (see also section 4.6). The PR effect also depends on the presence of evidence for missing information, i.e., noise has to be present leading to perceptual degradation.

Several paradigms have been used to measure this illusion, and to explore its bottom-up and top-down bases. Within the discrimination paradigm, for example, a second stimulus type is introduced by superimposing an extraneous noise and asking subjects to determine whether the stimulus is intact (noise superimposed) or not (noise replaced). The magnitude of the PR effect is reflected in the ability of subjects to make these discrimination judgments. These studies have shown that acoustic properties of the replacement sound (especially its psychoacoustic match to the speech it replaced) strongly affect the illusion, which indicates that the amount of bottom-up evidence has an important influence on the PR effect.

Another paradigm within which the PR effect is studied has been known as the multiple restoration paradigm (Bashford and Warren, 1987a). Here, speech is alternated periodically with noise, and subjects either adjust the alternation rate to find the largest gap duration permitting perceptual continuity, or are presented with speech interrupted at a fixed rate, and intelligibility is measured with and without interpolated noise in the periodic gaps. An important finding is that in order to preserve intelligibility, the minimum acoustic duration of the noise-interrupted speech segments should be around 40 ms, and the length of the noise-filled gaps at most 200 ms (see also section 5.3.4.3).

5.6.5.1 Lexical context

The PR effect also depends on listener-based factors, such as the amount of lexical activation of the tested word. These effects are particularly interesting in the context of lexical influences, and therefore the autonomy vs. the interactivity of the underlying processes leading to the PR effect which has been shown to be influenced by both previous and subsequent (lexical and/or sentential) context information. A variable that has been shown to influence the PR effect is, for instance, lexical status. The PR effect is stronger for words than nonwords, i.e., listener's ability to determine whether phonemes are replaced by noise or are noise-superimposed is worse in words than in nonwords (Samuel, 1981a, 1996, but not found in Samuel, 1987).

Nevertheless, for word-like nonwords (words that are consistent with the phonotactic regularities within a certain language), the PR effect also occurs, suggesting that low-level statistical regularities at a prelexical level again might play a role.

Also, word frequency effects have been found, such that stronger PR effects occur in high frequency words. This could result from faster lexical access, because more efficiently encoded or more stable representations exist for high-frequency words. However, the neighborhood density as well as the probability of phonotactic patterns within a word also plays an important role. For instance, Vitevitch and Luce (1998) have found that there exists a positive correlation between phonotactic probability and neighborhood density. This is because high-probability patterns are high in probability precisely because there are many words sharing the component segments.

If stimuli are used in which this correlation does not exist (for instance, sparse neighborhood density words with high-probability patterns, and dense neighborhood density words with low-probability patterns), a discrepancy is found. Stimuli with high-probability phonotactic patterns are responded to more quickly (Vitevitch et al., 1997), whereas stimuli in high-density neigborhoods are responded to more slowly (Luce and Pisoni, 1998). According to Vitevitch and Luce (1998), this reflects two different levels of processing:

(1) facilitatory probabilistic phonotactics reflects differences among activation levels of sublexical units, whereas
(2) effects of similarity neighborhoods arise from competition among lexical representations.

Results obtained with stimuli in which these two factors were separately manipulated were consistent with this interpretation (Vitevitch and Luce, 1998).

Regarding the phoneme restoration effect, Samuel (1987) found that lexical neighborhood effects influenced the PR effect such that with a high neighborhood density the PR effect is reduced, probably because lexical access slows down due to competitive effects (Samuel, 1987). Consistent with this finding is that the PR effect is also influenced by sentence context. Sentential predictability increases the PR effect, because of faster lexical access (e.g., Bashford and Warren, 1987a). It seems therefore that *phonemic restorations are influenced by prelexical as well as lexical factors.*

5.6.5.2 *Adaptation to perceptually restored phonemes*

One particular finding that has been claimed to provide strong evidence for interaction comes from a study performed by Samuel (1997). Using a selective adaptation paradigm, Samuel showed that words in which a given phoneme (/b/ or /d/) was replaced with noise led to a shift in the identification of stimuli on a /bɪ/-/dɪ/ continuum relative to the pre-adaptation baseline. No such effect was found when the phonemes were replaced by silence. The adaptation effect produced by the (presumably) restored phonemes was similar to the effect produced by words containing the intact /b/ or /d/ phonemes, i.e., there were more /b/ decisions when each adapting word contained a /d/ (and no /b/'s) than when each word contained a /b/ (and no /d/'s). Adaptation in the intact condition was not influenced by lexical factors. Samuel therefore argued that the effects had a perceptual locus.

However, McQueen, Cutler and Norris (1999) tried to replicate these findings (in Dutch), but did not find adaptation effects for the noise-replaced phonemes, whereas they did find reliable adaptation effects for the intact words as well as robust restoration effects for the noise-replaced conditions. Based on these findings, no conclusions regarding the presence of an adaptation effect or its possible locus can be derived, since adaptation can operate at several levels (see section 5.3.1.2).

Many of the described effects can also be "explained" by referring to the idea that there is a distinction between postlexical phonologically-based phonemic decisions and acoustically-derived phonetically-based decisions. The PR effect would then be based on phonologically-based codes. In this case, there is no need to suppose that lexical feedback has taken place leading to the perceptual restoration of the missing phoneme, since the missing phoneme is part of the activated lexical item.

Finally, it is important to realize that the study of the PR effect is complicated by the fact that, when natural stimuli are used, it is difficult to isolate and remove the acoustic information signaling the presence of an individual phoneme. This is due to coarticulation effects. Earlier findings already demonstrated that we are very sensitive to this subphonemic, featural information that is present in the context within which the phoneme occurs. This contextual, acoustic-phonetic information may therefore constitute a sufficient source of bottom-up information for making phonemic decisions and to activate and constrain the set of lexical candidates.

5.6.5.3 *Relation to Auditory Scene Analysis*

It is interesting that the effect strongly depends on the acoustic match between stimulus and phoneme. This is completely consistent with the findings that are generally found when studying the "continuity illusion" (chapter 4, section 4.6). Thus, it is important that the input consists of sufficient information that is consistent with the activated hypotheses or, perhaps more importantly, that does not provide mismatching information to reject the hypotheses. These hypotheses do not need to be lexical in nature, since prelexical sensitivity to phonotactic regularities could also lead to the sustained activation, or generation, of lexical candidates. Processing prelexical acoustic-phonetic features is facilitated due to the encoding of these low-level statistical properties.

Another point that should be mentioned here, is that in analyzing the results on the PR effect, the phoneme-replacing noise is taken as a whole sound. Though this is true based on our knowledge of the stimulus input, this is probably not how our auditory system deals with the noise burst. In chapter 4, it was seen that our auditory system performs auditory scene analysis. When the noise replaces a phoneme in a speech stream, there is good reason for the auditory system to assume that the speech stream continues behind the noise unless there is evidence to the contrary, such as when a silence is present. Because the global properties of the noise do not correspond to the global properties of a speech signal, in combination with the discontinuity that results from the sudden presence of the noise burst, the noise burst is ascribed to another environmental source. According to the old-plus-new heuristic, there is no reason to interpret the presence of another source as providing evidence for the absence of a previous source. The interpretation that the masked phoneme is absent would require the unlikely assumption that the maskee was switched off precisely when the masker was switched on. Therefore, (at least) two auditory streams are being formed when the noise is present.

As has been mentioned in chapter 4, between-stream temporal judgments are more difficult to make. This might explain why subjects are unable to determine where in the speech stream the noise has occurred, since the noise is not part of the speech stream.

Furthermore, it has also been seen that the so-called schemas[84] seem to take what they need from the sound mixture. The fact that the PR effect depends on the acoustic match between the noise and the missing phoneme, strongly indicates that, based on bottom-up information, it is possible to sustain the hypotheses that are activated, since the information they represent *is* actually present in the input.[85] Despite the fact that the information is ambiguous such that it can not be uniquely identified, there is no bottom-up mismatch information. Therefore, it is difficult to discriminate between the actual presence of a phoneme on which noise is superimposed, and the noise-replaced phoneme. Regarding the neural activation pattern resulting from the sensory evidence that is represented in the speech stream, and where the schema stands for, there is (hardly) no difference.

Though it could be argued that the global qualities of the noise are changed in the absence of a speech sound (for instance due to possible residue formation), the results presented in chapter 4 indicate that such a strict partitioning of data does not occur when different sound classes are assigned to different streams. This would suggest that the discrimination judgment can not be reliably based on such global differences, consistent with the inability of subjects to make a reliable discrimination judgment. Nevertheless, there seems to be evidence that at least part of the energy in the replacement sound gets "used up" by the restoration process (Repp, 1992: 2 experiments; Warren et al., 1994), though the evidence is mixed (not found by Repp, 1992: 3 experiments).

To conclude, subjects perceive the interrupted speech as continuous, since this is the most likely interpretation given the evidence. In coming to this interpretation of the perceptual evidence, the auditory system requires (1) the intact evidence, (2) the schema, which can either represent abstract information, e.g., vocal tract dynamics or prelexical, probabilistic phonotactics, or more concrete information, such as specific lexical knowledge, and (3) the noise. It does not require explicit reconstruction of intermediate representations. According to this interpretation, "phoneme restoration", or more generally the "continuity illusion", is essentially a metaphor. It describes the system behavior at a level where perceptual awareness takes place, and therefore refers to the percept itself, without constraining the nature of the underlying perceptual processes.

---

[84] I will use the term schema, since this term is generally used within cognitive science, and within ASA. However, I want to stress that I would like to apply this term in a very abstract sense, i.e., not by referring to a certain phoneme within a particular context, but rather by a pattern of activation that is initiated by current stimulus input based on its match with previously learned stimulus patterns. This is related to for instance processes like reconstructive memory. Though this is not inconsistent with the way a schema has been defined, since a schema can be either concrete or abstract, I prefer to make it explicit that the way it is used here refers to very abstract schemas representing implicit general knowledge at a very low level of representation, typically related to stored (vocal tract) dynamics.

[85] Notice that it is only possible to sustain the hypotheses that are already active from bottom-up input. The ability to do this is related to the fact that the auditory system always takes previous context into account (i.e., the auditory system is a dynamic system: the evaluation of current input is determined by the current system state which necessarily depends on the previous state of the system, since activity patterns are continuously updated in the light of new information). It is not needed (nor desirable) to make immediate decisions regarding for instance the phoneme identity of the noise-replaced or –superimposed phoneme during the time the phoneme is actually present. All that is needed is that the evidence presented earlier is (not in)consistent with the currently active hypotheses and subsequent input. Therefore, despite the inability to unambiguously identify the missing phoneme at a certain instant in time, its compatibility within a broader scope ensures that the auditory system is able to perceptually reconstruct an internal representation of the current perceptual scene. There is always a delay between our perceptual awareness (related to consciousness) of previous input. Since it is generally impossible for our perceptual system to represent different *in*consistent representations of the whole perceptual scene at one instant in time, it is necessary to solve possible momentary ambiguities over a longer period. If there is a sufficient basis for activating a particular word, there is no advantage in adjusting the activation of the features and/or the phonemes which gave rise to the activation of the word, since these are already part of the activated word representation.

Therefore, based on the ability of a computational implementation to simulate the PR effect, it is not possible to decide in favor of either autonomous or interactive accounts. Again, the choice of architecture and representations, and the assumptions on the accessibility of intermediate perceptual products, determine whether or not the effect needs to be "explained" by incorporating lexical feedback. The reported findings strongly indicate, however, that the PR effect reflects the "interactive" nature and redundancy of the underlying *representations* rather than the underlying processing between the postulated levels of representations.

5.7 Summary

Within psycholinguistic models of spoken word recognition, there are some general properties of the processing environment on which there is fairly broad consensus. For example, all psycholinguistic models that try to account for the dynamics of speech recognition, consider the process of lexical access as one in which multiple lexical candidates are simultaneously active. These can be seen as multiple hypotheses serving as output to higher levels of processing, which is consistent with serial models that allow multiple outputs like Shortlist and Cohort, as well as with interactive models like Trace (see also Boland & Cutler, 1996). So, on each level of processing, possible ambiguities are being preserved allowing for delayed commitment to a single hypothesis when needed, i.e., *the impact of a stretch of acoustic-phonetic input is not static.* This is reflected in retroactive effects that demonstrate the time delays between perceptual awareness and bottom-up processing.

The *activation metaphor* is considered to be the appropriate one for representing the goodness of fit between sensory inputs and stored lexical form representations, such that lexical items are activated according to their degree of similarity with the acoustic-phonetic input. As time passes by, the *amount of lexical activation builds up gradually.* All models specify that activation on all levels of processing is *continuously* updated in the light of newly incoming information, which means that *partial information is made available immediately.* Activated lexical items remain active as long as the sensory evidence is not inconsistent with the activated hypotheses. In the case of missing evidence, there has to be evidence that information could be "missing" - or actually, less accurately identified - through masking or other perceptual degradation phenomena, which is consistent with the findings reported in chapter 4 concerning the continuity illusion.

Lexical activation is *very sensitive to bottom-up mismatch information.* In the existing models this is realized through direct inhibitory bottom-up connections (Cohort and Shortlist). In contrast, models such as Trace indirectly correct for mismatching information through lateral inhibition at the lexical level. Here, the more active candidate, i.e., the one most consistent with all the available information, reduces the activation of less active candidates more than vice versa (contrast-enhancing). Experimental evidence indicates that the latter procedure depends too much on the lexical status of incoming stimuli, and can not account for the observed sensitivity to mismatching bottom-up information.

A *distinction* is made between *generation and selection processes,* and between *lexical access and word recognition.* Generation processes of lexical candidates are autonomous, and are based on the evaluation of acoustic information. This process is not influenced by a preceding biasing sentence context. Subsequent context can influence the final perceptual interpretation of earlier presented ambiguous, acoustic information (that led to a sufficient amount of activation for lexical access to occur). Given the often not optimal listening conditions, this allows the system to be flexible. Nevertheless, there are temporal limits within which decisions are made. In general, contextual effects reflect selection processes. Contextual improbability can have an inhibitory effect, and contextual probability has a facilitatory effect. This facilitation can be achieved (1) by lowering the threshold level (criterion-shift), which results in a perceptual *bias,* such that less sensory information is needed to reach a level sufficient for recognition, and/or (2) by increasing the activation level of lexical candidates that are activated from bottom-up input, thereby giving (independent) additional evidence. The latter allows for contextual information to pull activated lexical items further away from their competitors.

The selection *decision* is based on the *relationship between levels of activation*. This can be instantiated, for instance, (1) via a contrast-enhancing mechanism such as lateral inhibition, (2) via a best-one-wins mechanism, by comparing *relative* levels of activation and making a decision as soon as one candidate diverges enough from other active candidates, (3) by *absolute* criteria, i.e., when some internal deadline (in time, or in threshold level) is passed, with the possibility of criterion-shifts, especially when additional, consistent sources of information are present, (4) by a postperceptual decision process where the activation values are transformed into probabilities, thereby making the perceptual evaluation processes deterministic, and the decision process stochastic in nature, or (5) by allowing the perceptual processing mechanisms to be stochastic in nature during processing, such that at each instant in time, the evaluation of features takes place in informational terms, which directly leads to activation values that represent the likelihood of selection. In either case, the selection process results in competitive behavior of the recognition system. The existing models differ in the locus and nature of this competition process. Often, however, these distinctions can be traced back to the particular choice of architecture and representational units, which is also related to the level of description at which the system is described.

Chapter 6 *General Conclusions and Discussion*

In this chapter some issues will be discussed in the context of some recurring themes that were encountered throughout the different chapters. These include the following:

- The dynamic nature of representations at multiple time-scales: the ability to continuously track the stored dynamics, as represented by (abstract) features that evolve along the time dimension (section 6.1)
- Retro-active effects in perception: the difference between the time-scales at which conscious awareness takes place, and the process of perception that continuously responds to incoming stimuli (section 6.2)
- Learning and self-organization: the importance of the ability of our cognitive system to adapt its internal structure and/or internal state to the requirements of the environment through a self-organizing learning process (section 6.3)
- Sensory data versus information: the often encountered distinction between current stimulus input and stored knowledge, and their joint contribution in perception (section 6.4)
- System architecture: some general properties of the system architecture that seem to be essential based on the findings presented in this literature research (section 6.5)

6.1 The dynamic nature of representations at multiple time scales

Section 5.3 presented an overview of the experimental findings obtained within research aimed at specifying the underlying representations that are formed during the process of speech recognition. Many findings support the idea that *acoustic-phonetic features are extracted from the speech signal by means of acoustic cues*. These acoustic cues are interpreted such that they correspond to *subfeatures* that can be estimated within a time-span corresponding to about 40 ms, and that constitute a first level of abstraction of the speech signal (section 5.3.4.3). Off course, the ability to extract these features in arbitrary circumstances largely depends on the ability of groups of neurons to group the relevant signal components. This can be achieved through, for instance, the temporal encoding characteristics of neurons. For voiced speech, the grouping *emerges* from the sensitivity of neurons to the temporal fine structure of the speech signal, e.g., the synchronous firing of neurons with the period of the fundamental frequency, which reflects the pitch of the sound source.

Furthermore, the *features reflect the temporal dynamics characterizing the speech signal* that are *related to the dynamics of the underlying articulatory processes* (e.g., section 2.3, 2.4, 4.12, and 5.3.3). Examples are:

- The detection of *formants*: These correspond to frequency regions that are more amplified due to the filtering characteristics of the vocal tract (as described in section 2.1.3). The underlying neural processes that encode these formants depend on the acoustic sig-

nal properties, such as the presence of voiced components (not only harmonic patterns, but also amplitude modulation (AM) patterns) versus bands of noise bursts (section 3.2, and 3.9.2). In general, voiced speech is more noise robust by virtue of the more effective entrainment of BM regions in combination with the phase-locking capabilities of neurons in the auditory system (see also section 5.3.4.2).

- *Formant transitions*: These correspond to an abstract *pattern* of FM, that is no longer tied to the particular frequency region in which it occurs, and can be used to distinguish between consonants-in-context (e.g., sections 2.2.2.4, 2.4.2, 3.9.2.3, 4.9.3, 4.10, 5.3.1.2, and 5.3.3).
- The *integration of formant patterns*: This allows for the encoding of the *relative* pattern of change of different formants. This takes place at different levels of abstraction: an integrative, acoustic level, and a more categorical level that is probably related to a phonemic interpretation of the signal as speech (section 4.12, 5.3.1, and 5.3.2.2). Though the presence of pitch allows for the grouping of related signal components underlying one formant pattern, this level is not very sensitive to a consistency between the pitch underlying the different formant patterns that are integrated. Nevertheless, if the pitch underlying different formants is the same, they will tend to become perceptually more salient as a group.
- *Voice onset time*: VOT is related to the *(relative) time difference* between two auditory elements (as mentioned in section 2.2.2.4, and 3.9.2).

These *features are continuously integrated along the time dimension* such that *partial information obtained within relatively small frequency channels becomes immediately available* (section 5.3.2.3 and 5.3.3). It seems very plausible that the *syllabic unit provides a coherent source of segmental information* (section 5.3.4.5). The mean length of a syllable roughly corresponds to the length of temporal integration windows (~200 ms) at the level of the auditory cortex (AC), and forms a further abstraction away from the spectral detail that is present in the physical signal. The decreasing temporal resolution in the AC (mentioned in chapter 3) also indicates that this abstract representation of speech provides a better generalization that is relatively insensitive to acoustic variations. The relation between the syllable and morphological units of meaning strengthens the syllabic unity. Furthermore, in section 2.2.1.5, it was mentioned that phonological rules - related to articulatory processes - could be formalized much more efficiently within the context of a syllable. This way, subphonemic features can be used to distinguish between different syllables.

Based on the results on phonetic categorization tasks (e.g., section 5.3.2.2), it is not possible to determine whether phonemes are perceived categorically, or whether the features characterizing the phonemes-in-context - and therefore the syllable itself - are perceived categorically. In general, care should be taken when interpreting the evidence on phonemic decisions tasks, since this task requires subjects to explicitly focus attention on the sound structure of their language. It seems that this does not characterize normal language behavior, and additional knowledge resources are used to accomplish these tasks (section 5.3.2.4). This is also the case for tasks in which nonword stimuli are used (e.g., 5.3.2.3). It is important to study the process of speech recognition as much as possible within natural circumstances, using natural stimuli. Since the observed language behavior results from a process of learning, stored knowledge (representing environmental regularities) always influences subsequent processing. This can not be isolated when modeling the underlying processes in speech perception.

To conclude, the results concerning the nature of the representations that are formed during spoken word recognition are quite consistent with one another. They converge on many findings presented in the different chapters, specifically the importance of the dynamic properties of speech, and the ability and sensitivity of the perceptual system to accurately represent these temporal relations at different levels of abstraction.

6.2 Retro-active effects in perception

The retro-active effects in perceptual awareness reflect the relevance of the distinction that is made within psycholinguistics between the *recognition* of a lexical item and lexical *activation* (section 5.1 and 5.4). Lexical activation is based on *partial* recognition, without the unique identification of one particular word. Word recognition reflects the outcome of perceptual processing in that we only become aware of *one* single interpretation of simultaneously present auditory information that is consistent with all other available sources of information (either present in the environment or represented in the connectivity pattern between (groups of) neurons, which is the result of an ongoing learning process). During processing, there is no need for such decision-making when the input is (still) ambiguous.

Similarly, acoustic cues to the presence of features can be shared out whenever they are consistent with a certain feature. This is consistent with knowledge on auditory functioning, where the same acoustic cue is used to estimate different sources of information, sometimes by first transforming it into another code (see chapter 3). It is also consistent with the violation of the principle of exclusive allocation in ASA as far as the acoustic cues are concerned (section 4.4.1 and 4.10).

Advocates of bottom-up autonomous models therefore argue that acoustic analyses of the auditory input should not be influenced by processing based on other sources of information, but should preserve their autonomy and retain their input-specificity. Ambiguities should therefore be preserved (section 5.4.2). But, of course, for the purpose of making decisions (which ultimately are required to achieve word recognition), results of all available sources of information are continuously integrated and therefore they all exert their influence on the perceptual outcome (which is normally consistent with all available sources of information). Though this may lead to the recognition of a word based on less auditory information through the use of top-down knowledge,[86] this can only occur when sufficient bottom-up auditory information is present in the input leading to the activation of the presented word. Without this bottom-up evidence, i.e., without the generation of lexical candidates, recognition is not possible. Indeed, in exceedingly ambiguous situations (for instance, with very low signal-to-noise ratios in the order of –6 dB), human speech recognition deteriorates rapidly, because there is no sufficient basis for lexical activation. This is consistent with what has been shown regarding the continuity illusion (section 4.6), and the phoneme restoration effect (section 5.4.3.3 and 5.6.5).

6.3 Learning and self-organization

An important characteristic of the human speech recognition system is that it is a self-organizing system capable of coding the (behaviorally-relevant) regularities that are present in the environmental stimuli to which it is exposed (section 3.9, 5.5.4.2, 5.6). As a result, the functional organization within the brain reflects this environmental structure.

[86] Notice the distinction between the use of top-down knowledge and the use of top-down feedback. In autonomous models of spoken word recognition, top-down *knowledge* (as it is usually called) can aid directly in processing, since such knowledge is implicitly encoded within long-term memory (LTM). Memory traces become reactivated when the appropriate cues are presented. Once activated, knowledge resulting from previous experiences, also becomes available, a process known as reconstructive memory. However, it is claimed that this does not affect the bottom-up generation of lexical candidates which is based on a massively parallel process where perceptual encoding takes place based on the current stimulus input, and ambiguities remain unresolved. Since these processes are sensitive to lower-level statistical regularities of the environment, they also represent a form of "top-down" knowledge in their connectivity pattern. Incorporation of top-down *feedback* alters these early processes of perceptual analysis.

Processing models should be evaluated against the so-called "self-organization critique": an internal analysis of the model from the viewpoint of whether its processing mechanisms could, in principle, develop or be learned. A model which cannot self-organize must be using certain mechanisms that are physically incorrect. It has been argued by Grossberg (1991) that the nodal units (representing phoneme and word representations separately) and the internodal interactions postulated in McClelland and Rumelhart's Trace model are seriously challenged by this critique. Grossberg also stresses the importance of defining a processing substrate that can represent the learned units of subject's internal lexicon *before, during,* or *after* they are learned.

Unfortunately, the existing psycholinguistic computational models of speech recognition are hardly motivated by such concerns. The justification and falsification of a model occurs (mainly) on the basis of its ability to simulate behavioral data, not on its consistency with neurobiological knowledge regarding the processing substrate. Such a description of how the system behaves can be considered as a functional system specification. A functionalist approach is characterized by the concept of *multiple realization,* which implies that the *underlying mechanisms are actually irrelevant* at this level of description (Kim, 1996). If the purpose is to describe the internal processing within the human speech recognition architecture, it is impossible to distinguish between models such as Trace, Shortlist and Cohort *based on* their ability to simulate behavioral data. The different models can simulate the same pattern of results, but *the nature of the explanations that are given are the direct consequence of the choosen computational processing architecture,* and therefore *not necessarily the human processing architecture.*

For instance, localist connectionist networks such as Trace and the lexical competition process in Shortlist make quite strong and explicit assumptions about the structure of the internal representations and the processes that manipulate them. They therefore do not place much emphasis on how the learning system would evolve such that it leads to these presupposed internal representations. Emphasizing form and structure rather than learning and adaptation leads to different "explanations" of the observed behavior. Both explanations describe dynamic aspects of processing, but at different levels of descriptions.

6.4 Sensory data versus information

Usually, top-down knowledge refers to previously learned knowledge that is stored in long-term memory thereby influencing the interpretation of incoming sensory evidence. Within the psycholinguistic literature, and within some approaches to perception this is sometimes interpreted as referring to a top-down flow of information where lower level representations are altered in the light of higher order knowledge (see section 5.6). Partly, using this terminology is due to the choice of the architecture or theoretical framework within which experimental results are being simulated and/or explained.

In general, perception involves a reconstruction of internally stored representations. Within neurobiological approaches and approaches that try to account for the *microcognition,* this is usually characterized as a distributed pattern of activity in the brain. The combination of memory traces that are activated by the sensory data, mediated by the consistency that exists between stored associations reflecting (meaning-related) world knowledge, eventually leads to a coherent overall interpretation of the environmental scene.

This process can also be described, at a functional level, as a constraint-satisfaction process, where the perceptual system "searches" for an optimal interpretation of the sensory data. This *functional characterization* of the perceptual system is reflected in the interactive Trace model (section 5.2.2), the selection stage in Shortlist (section 5.2.3), and the competition and collaboration of cues as described by Bregman (section 4.4).

The distributed Cohort model, on the other hand, seems to become closer in characterizing the *underlying mechanisms*, since it places more emphasis on learning and self-organization leading to stable organizations within an attractor network (section 5.2.4 and 5.3.3). Within localist connectionist networks these are replaced by semantically-transparent symbols, whereas distributed, more dynamic, attractor networks describe the emergence of these stable states subsymbolically. In either case, the final perceptual outcome is always the result of interpreting the data in the light of stored knowledge.

6.4.1 Statistical knowledge and bias effects

Nevertheless, early in processing, the *physical*, acoustic cues that are related to some of the characteristics of natural sources are processed automatically and accurately, and reflect what is present in the data. The fine-tuning of these mechanisms occurs early in development, and because of their high degree of generality, they remain relatively static. The ability for these mechanisms to develop is supported by the general response and learning characteristics of neurons and the tonotopic organization throughout the auditory system.

The more abstract cues that are *derived* from the physical signal, such as acoustic-phonetic features, are also accurately processed such that at lower levels of processing, possible ambiguities as to how they should be interpreted within the more global picture are retained. Nevertheless, for these signal properties, processing is influenced as a result of associative learning. The influence of previous exposure with auditory stimuli in the environment leads to an (implicit) encoding of its statistical properties. As a result, these statistical regularities allow for more efficient hypothesis-generations at higher levels of processing (and therefore an adequate reduction of the possible search space).

This is reflected in the *sensitivity to regularities in the pattern of vocal effects resulting from vocal tract dynamics*, as described in the chapter on ASA (section 4.12). This results in the ability to *group* signal components that might have arisen from one vocal source before more specific recognition schemas come into play. The so-called *prelexical sensitivity to phonotactic constraints based on statistical regularities existing within the sound structure of the language*, as described within psycholinguistics, seems to reflect the same principle (section 5.5.4.2). Within bottom-up models of spoken word recognition, this sensitivity is argued to reflect the autonomy of earlier generation stages that result from a feature evaluation process (section 5.4.1). This can be interpreted as a prelexical filtering stage that continuously feeds processing at higher levels as soon as new sources of partial information are available. At this stage, statistically-based trading relations between cues and/or features already exert their influence on processing in how they activate stored lexical representations through the connections that exist between the two. This roughly corresponds to a *feature integration* stage. Nevertheless, they are not constrained by other "knowledge resources" stored in long-term memory.

For a large part, the *bias effects* that have been found in psycholinguistic studies on word recognition also reflect the effects of (perceptual) learning (e.g., the Ganong effect, the "compensation for coarticulation" effect, and the "phoneme restoration" effect, described in section 5.6). Bias effects can therefore be explained in terms of a prelexical sensitivity to regularities within the language, priming (pre-activation) of lexical items, the presence of lexical representations (corresponding to stable states for which the perceptual system exhibits fast settling behavior) for words as opposed to nonwords, and more stable and more efficiently encoded representations for high-frequency as opposed to low-frequency words thereby requiring less information in order to be recognized.

6.4.2 Context-sensitivity and perceptual discreteness

The concept of information plays a very important role, especially at higher levels of processing, where behaviorally-relevant *information-bearing parameters* (IBPs) allow for making meaningful distinctions. It is difficult to define information in perceptual processing unambiguously, since it can not be quantified by *entropy*. This is due to its subjective component, i.e., its meaning dependency: the meaning or informational content of a particular source of information differs among subjects and is related to the personal history of experiences of a particular subject which changes continuously. Nevertheless, it can be qualified as being related to *discrimination pressure*, and it is therefore analogous to an increase in *information gain*, i.e., a reduction of entropy. The higher the discrimination pressure, the more specific the IBPs, and therefore also the more perceptually discrete. This results in more categorical behavior (section 3.9.1).

In a sense, this corresponds to the discrete linguistic-phonetic features that are contrastive within a certain language, which are argued to allow for the distinction between phonemes (described in section 2.2.1.4). However, an important characteristic of IBPs is their multidimensionality, i.e., their context-sensitivity, which makes them act more like true feature-detectors. This means that they only become activated in the context in which they are relevant. In order to be able to do this, they must be strongly inhibited when contextual information is inconsistent with the information they represent.

In contrast with this stored context-sensitivity, linguists postulate *phonological rules* to account for the context-sensitivity of features (section 2.2.1.4). These are needed to explain the variance in the distribution of linguistic-phonetic features signaling the presence of a particular phoneme as a result of articulatory processes. By applying these phonological rules the underlying phonemes can be derived. Some psychologists have also argued that the task of the perceptual system is to *restructure* the cues in order to compensate for coarticulation effects (as reflected in, for instance, the Trace model, see also section 5.6.4).

However, within a dynamic representation that encompasses the context, there is no need for such a compensation processes (see also section 2.3, and 5.3.4.5). Nevertheless, due to the higher degree of similarity between allophones of a certain phoneme class than between allophones of another class, it is possible that more abstract knowledge regarding phoneme classes is represented by activity in a particular region of the brain (comparable to the reasoning in 3.9.2.1 regarding vowel recognition). This way, phonemes can be represented more or less categorically, but for their perception during speech recognition this is less relevant.

6.4.3 The influence of attention

Notice also that many experimental effects seem to depend on the locus of attentional focus. Selective attention processes are often considered to result in facilitatory effects in processing relevant information and/or inhibitory effects on processing irrelevant information, which leads to contrast-enhancing behavior, i.e., perceptually more discrete perception. During normal speech recognition, the purpose is not to recognize words, but to derive *meaning* from the relations between recognized words. Meaning is not present in the individual words, but is related to world knowledge, i.e., the subjective internal representations that reflect the structure that is perceived to be present in the environment. Within tasks on word recognition, attention is focused on the word level. Within phonemic decision tasks, attention is focused on a phonetically-related level. It may well be that this attentional focus leads to the observed competitive behavior due to stronger contrast enhancing effects (e.g., categorical perception). Normally, information processing at lower levels of representation

seems to be more graded, certainly not all-or-none. This is reflected in, for instance, psycholinguistic bottom-up models that are characterized by the absence of inhibitory mechanisms at lower levels of representation (e.g., Shortlist and Cohort).

Therefore, the recognition system is required not to be confined to using sections of the speech stream to identify only a single word at a time. This is in line with a characterization of speech recognition aimed at deriving meaning from recognized words within a broader temporal scope (see also sections 5.3.4.7, 5.4.2, and 5.4.3.3). Furthermore, it can be easily instantiated in systems where supporting sensory evidence leads to a gradual build-up of activation (as has been argued in sections 4.5.4 and 5.4.3), in combination with a high degree of sensitivity to inconsistent sources of information (consistent with the results described in, for instance, section 5.3.2.2, 5.3.3.1, and 5.5).

6.4.4 Learning and competition effects

Within some psycholinguistic models, bottom-up inhibitory connections are included to account for the observed sensitivity to bottom-up generated inconsistent contextual information (section 5.5.4.1). This corresponds to some extent to neural processes such as *disinhibition*, where a neuron only responds to the stimulus information for which it stands, and is inhibited such that it does not fire in the presence of contradictive information. This differs from lateral inhibition, since this results in contrast-enhancing behavior, but does not directly prevent the neuron (or group of neurons) from becoming active. Again, functionally, these different descriptions may result in the same kind of categorical behavior that has been observed experimentally, but they differ with reference to the underlying mechanisms that are needed to account for this behavior. They also differ in the predictions they make regarding, for instance, lexical competition effects. It has been argued that explanations of these competition effects based (solely) on lateral inhibition at the lexical level are less adequate.

However, concepts such as information, adaptive learning, and self-organization may again be relevant in order to, perhaps more accurately, explain the competition effects that have been found within psycholinguistic experiments. This will be reflected in the following speculations that are based on knowledge regarding learning and memory retrieval, where it is argued that *competitive learning leads to competitive effects during perception.*

During learning and development (which of course never stops, but can be described at a larger time-scale compared to the "on-line" perceptual processing), the statistical properties between cues and features are learned. If a certain cue acts as a reliable predictor for the presence of a feature, its association strength becomes stronger. However, in the case of high density neighborhoods, the predictability of one particular lexical item based on the features characterizing the item is less than in low-density neighborhoods (see also section 5.5.2). This is because there is more ambiguity (or uncertainty).

In memory retrieval, the activation of a particular item depends on

(1) the amount of perceptual (bottom-up) evidence,
(2) the association strength between cues present in the input and the lexical item, and
(3) the record strength of the lexical item.

The record (or memory) strength is related to the statistical nature of the lexical item. High-frequency words have stronger record strengths than low-frequency words. Usually, this leads to faster memory retrieval and therefore faster recognition (according to a *power law of learning*[87]). However, when the input is increasingly ambiguous, due to the presence of noise and/or due to the presence of multiple consistent lexical candidates having a high degree of

[87] See any book on learning and memory (e.g., Anderson, 1995).

phonological overlap, the predictability of the evidence characterizing the lexical item is weaker, and hence the growth of lexical activation is also impaired. This might be ascribable to the weighting of evidence resulting from competitive learning, which is reflected in the association strengths .

If an organism is to learn the statistical properties of the environment, it is not only sensitive to the frequency of occurrence of one particular lexical item, but also to the predictability of the features that *characterize* the lexical item. This predictability is influenced by the pattern of occurrence of *other* lexical items sharing the same set of features, and therefore such a *relative evaluation of evidence* is expected to be reflected *directly* in the build-up of activation during perception (i.e., memory retrieval, or reconstructive memory). Therefore, during perception, the "same" amount of perceptual evidence can lead to less activation of a particular lexical item when it is to be "divided" among multiple lexical items. Contrary, when it is assumed that a certain amount of perceptual evidence leads to the same amount of activation of a lexical hypothesis, independent of the predictability of the evidence, inhibition is required to correct for this, in my opinion incorrect, assumption, because the activation builds up too rapidly (see, for instance, section 5.5).

Also, cues that add to the ability to *discriminate* between features characterizing particular lexical items, have a higher degree of predictability, i.e., are more reliable, or less ambiguous, and therefore more informative (related to the IBPs described in section 3.9.1, and to the distinctive features described in section 2.2.1.4). Therefore, the association strength of such features to the lexical items they characterize is relatively stronger, and they exert a relatively greater influence upon the activation growth of the lexical item. *If the amount of activation is supposed to reflect the amount of certainty of a certain hypothesis, decreasing the ambiguity must necessarily lead to an increase in the build-up of activation.*

In Cohort, activation builds up for each lexical hypothesis independent of the presence of other hypotheses. A selection mechanism is used after the evaluation and integration of features to choose among alternatives. It is at this stage that the likelihood that the subject selects a certain response depends on the presence of other activated hypotheses. There is no need to suppose such dissociation between activation and likelihood of selection. Response selection of neurons is most important, *not* the responses of subjects. Activation and selection are therefore "parallel" processes: the "decision" of a (group of) neuron(s) to respond to certain input is directly influenced by (relative) likelihoods. If two hypotheses are similar in some aspect they share *mutual information*. Hence, they are *not* completely independent. Such mutual information does not provide an effective means to discriminate between the two, and therefore does not provide the most useful source of information. In fact, it does not provide information, because it does not reduce the entropy, i.e., there is no information gain. If ambiguity is increased by presenting lexical items in noise, the similarity is increased, and discriminative features are even more important. When these can no longer provide a sufficient source of information, because of the reduced ability to estimate the feature (values), this should lead to a slower growth of activation, because there is no single pattern of activation that diverges enough from the other patterns of activation to lead to settling behavior.

In general, it can therefore be stated that neurons fire such as to optimize the amount of information gain, i.e., a high rate of information transfer is preserved from input patterns to output pattern. The rate of information transfer from input to output can be found directly from the mutual information between input and output. This results in a reduction of the entropy, since there is less uncertainty (i.e., a decrease in ambiguity) when a certain neuron becomes active, or a group of neurons remains active. Accordingly, there is a large variance

of the output activation for a given set of input neurons,[88] resulting in adequate generalizations in combination with the required input-specificity. This way, the survival of organisms whose perceptual systems are best adapted to their environment is strengthened.

An analoguous line of reasoning *describes* competitive "inhibition" as resulting from a diffusion of activation ("*energy*") during processing: when more hypotheses are consistent with the input, the evidence is weighted according to the already obtained amount of activation based on previous input and the amount of evidence provided by current input. This leads to a slower activation function and lower final levels of activation in the context of multiple consistent hypotheses. Initially, there is not enough information to make a sufficient enough distinction and activation builds up slowly because of this high degree of ambiguity. As time passes by, less candidates can account for the input presented so far, and a faster build-up of activation for the different candidates is possible (because the energy is divided between less candidates). When even more time has passed by, more of the ambiguity is resolved and less information is added. The activation level saturates. This could partly "explain" the general form of the activation function that is often used, i.e., a non-linear sigmoid function. The slope of this sigmoid function is determined by the *relative* weighting of activation, and is therefore directly related to the number of possible candidates that are trying to account for the incoming sensory evidence. The reduction of the activation level due to competition is therefore attributable to this slower growth of the activation level which *is* affected by the presence of other candidates but not through the type of inhibition proposed within psycholinguistic models[89] (i.e., lateral inhibition between (phonemes and) words

When perceptual processing is considered to be a new *learning instance*, the descriptions in terms of a diffusion of energy during processing, and in terms of competitive learning mechanisms can actually be considered to refer to the same processes. This blurs the distinction that is often made between learning and perception. When considering the possible underlying mechanisms, the distinction is, in a sense, irrelevant. Nevertheless, these processes can still be described at different time-scales, and the used terminology is therefore useful for distinguishing between these levels of description. After a period of learning, the set of association weights have become more or less stabilized, and the resulting behavior is therefore more influenced by the processing resulting from previous learning than by the current learning instance, i.e., the current stimulus input.

[88] This could be implemented, for instance, in a layered network with feedforward connections and (not necessarily) linear activation functions, in which a simple Hebbian type of learning rule builds layers of units that have progressively more sophisticated feature-analyzing properties. Cells are developed that maximize the variance of their output values, and therefore perform Principial Component Analysis (PCA), a sort of feature extraction, on their inputs, and preserve maximum information about the input activities. The algorithm can be described as follows: Given input patterns from an n-dimensional vector space, find some subset $m<n$ of the n variates that accounts for as much of the data's variability (variance) as possible

[89] Notice however, that the spreading of activation (and the relative weighting associated with it) is, in principle, only true for activation that results from the *same* source of information (i.e., both stimuli are auditory stimuli). Facilitatory effects are typically found for (cross-modal) semantic priming experiments, which is based on another source of information: activation resulting from the presentation of visual stimuli adds energy, and in general, activation resulting from the pre-activation of semantic associates also results in facilitatory priming.

6.5 System architecture

Based on the findings and conclusions described in this chapter and in the previous chapters, the system architecture of a computational speech recognition system should exhibit the following properties:

- *Redundancy of representations*: The system can profit from the *redundancy* that results from using *multiple cues*, based on possibly *multiple signal representations* of the same signal component (the same acoustic cue can be used multiple times, whenever needed for determining the presence of a certain feature). This property makes the system more noise robust as it is less dependent on a limited set of informational sources of whom none can be trusted completely or is available in all circumstances.
- *Autonomy/Modularity*: The processing of different cues is as *autonomous* as possible. At the same time, *independent* and *consistent* estimates strengthen one another at an *integration* stage.
- *Scaleability*: The system should be *scaleable*, such that it is capable of dealing with more complex problems in a natural way, and to use more cues than it is originally designed to deal with, without completely redesigning the system architecture (related to the previous point of autonomy/modularity).
- *Assimilation*: The system should be able to *assimilate*, i.e., to apply its "rules" to arbitrary sounds, if and only if the input requires the activation of these rules to deal with the presented input.
- *Multiple hypotheses*: The system has the ability to deal with *ambiguities* by allowing the *activation of multiple hypotheses*, and by leaving the possibility of delayed commitment to final hypotheses. So, sensory data is integrated within a *global temporal scope*, such that it does not depend on temporary ambiguities.
- *Context-sensitivity*: Context is always taken into account. This is related to the fact that *the AS is a dynamic system* where the mapping takes place against *dynamic representations*. Therefore, the system should be able to deal with *interdependencies* (trading-relations) between different cues. This can be achieved, for instance, in the form of *competition* (inhibition) and *collaboration* (excitation), and by allowing a *propagation of constraints* (functionally analagous with a constraint-satisfaction process).
- *Adaptation/Flexibility*: Properties such as assimilation and context-sensitivity require an *adaptive, flexible* system architecture, that includes the ability of *learning*, and the ability to *represent appropriate abstractions/generalizations*. As a result, the *perceptual evidence is always interpreted in the context of stored information*, i.e., it is influenced by previous experience, and is able to internally represent and become sensitive to the structure that is present in the environment.
- *Self-organization*: Ideally, the system should be able to self-organize, since this naturally reflects most of the above mentioned properties. However, for a computational implementations, this can be approached by allowing the system to incorporate processing principles and representational transformations by taking *advantage of knowledge about human auditory information processing*, and in particular, speech recognition. This also implies that the *coupling* between the preprocessing stages that form the input to higher levels of abstraction is as optimal as possible.

Though this list of properties of the system architecture is not complete nor explicitly enough formalized, the findings presented in this literature overview suggest that these are *essential* properties. Therefore, it at least provides a check-list for the possible architectures, methodologies and techniques that are worth considering when the purpose is to build ASR systems with a level of performance that is comparable to that of humans.

Appendix A *Signal Detection Theory*

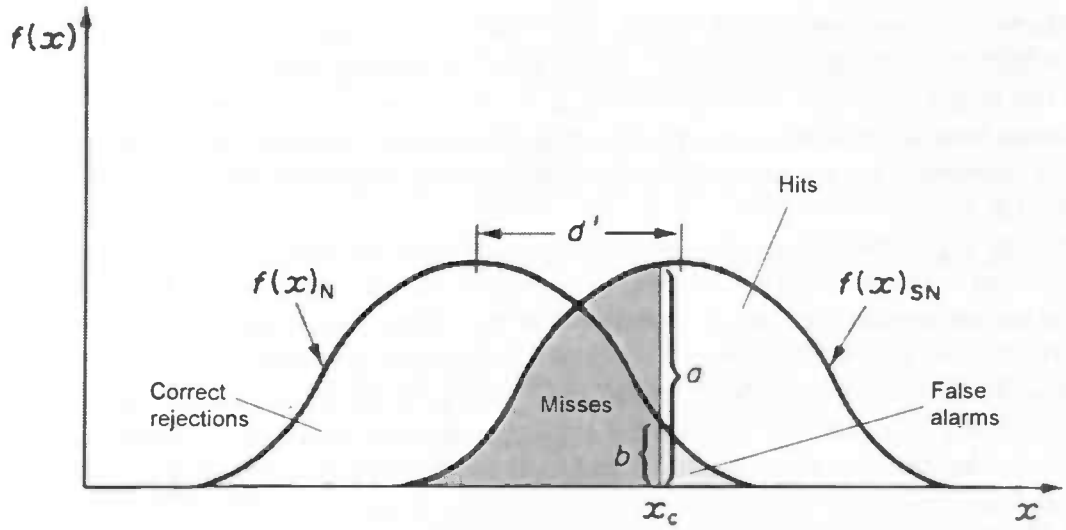


Figure A.1 The two distributions (probability density functions) which occur in response to "noise" alone [$f(x)_N$], and noise-plus-signal [$f(x)_{SN}$]. The separation between the means of the two curves, denoted by d', gives a measure of the discriminability of the signal. The criterion value is labeled x_c, and the ratio and the heights of the curves at x_c, namely a/b, gives the likelihood ratio, β.

Signal detection theory (SDT) can be applied to *discrimination* as well as *detection*. SDT asserts that decisions about perception are fundamentally probabilistic. Given the large variability that is typically observed in perceptual experiments, this approach seems inherently reasonable. The stochastic character in perceptual experiments is modeled as noise, either as part of the input received by the decision-making system or as part of the operation of the system itself. Such a model fits well with the observed behavior of sensory neurons, exhibiting spontaneous random firings in the absence of external stimuli and a noisy response when a stimulus is present. SDT can therefore serve as an interface between psychophysical observations and (neural) models of perception. SDT provides a framework that can unify the results of different experimental methods as it provides a characterization of an observer detecting a signal against a background of noise, according to which the behavior of the observer can be desribed by two statistics (see figure A.1):

(1) *Sensitivity* measure d' (d-prime), referring to the distance between the signal and noise distributions, or the ratio of the size of the response to the signal to the size of the standard deviation in the response to the noise. The further apart the distributions are, the more sensitive the observer is, and the easier it will be to discriminate between the presence of a signal and noise.
It is assumed that there is an internal coordinate, x, that the listener uses for a decision (e.g., neural firing rate, or a synchrony index). There is a distribution representing excitation on different trials, and the presence of the signal displaces the distribution along the internal coordinate (which is assumed to be the only effect of the signal, so that the standard deviations of f_N and f_{SN} are the same).

(2) *Bias* measure β, or the *likelihood ratio*, referring to the ratio of the probability of correctly responding to the signal (a) and the probability of responding to the noise (b) at the criterion value x_c. Ideally, a subject should adopt a criterion that maximizes the *hit rate* while minimizing the *false-alarm rate*. However, in setting the response threshold an observer may have a bias to respond "signal" or "noise", which can be determined by calculating the value of parameter β.

An example of an experimental condition in which SDT can be applied is a *Two-Alternative Forced Choice* (2AFC) task, e.g., a subject is presented with two observation intervals in succession, where only one contains the signal. After the two stimulus intervals, the subject has to decide which one contained the signal. Given that the subject chooses the interval associated with the larger response, the probability of a correct response is the probability that a sample chosen from distribution f_{SN} is greater than the sample chosen from f_N (i.e., $\Delta x = x_{SN} - x_N > 0$). The important parameter to be estimated from the experimental data from a 2AFC experiment is d'.

Another experiment is the *Yes-No experiment*, which involves only a single observation interval that may contain a signal (SN) or it may be the noise alone (N). The subject has to decide whether the signal was present (Yes) or not (No). This experiment involves the subject's criterion as well as the sensory character of the task. For example, from a subject's Yes-response it can not be concluded whether the signal is much stronger than the noise (i.e., the discriminability is higher) or whether the subject is eager to say "Yes", or reluctant to say "No". SDT enables the experimenter to distinguish between decisions based upon sensory characteristics and response tendencies by estimating the parameters d' and β. Measures of sensitivity and bias can be derived from the observer's *hit rate* (e.g., correct Yes-responses to signal) and *false-alarm rate* (e.g., Yes-response to noise, in absence of signal), which are *assumed* to be *independent* of one another. [90]

[90] Applying SDT to a set of data carries with it a set of theoretical assumptions about the processes being studied. The claim that changes in measured sensitivity are a direct reflection of changes in the sensitivity of some early perceptual process is only true under the specific set of assumptions made by SDT. SDT statistics only provide a description of the observer's *behavior*, so a measured increase in sensititivity only tells us that the observer is behaving as if the signal and noise distributions had moved further apart in a system that makes response decisions in exactly the way proposed by SDT. If these assumptions are not met, then a change in d' need not imply that the observers's perceptual processes have changed in any way at all. In more complex systems, such changes in apparent sensitivity can also come about through changes in the decision processes or biases involved in response selection. So, an increase in d' of an encoding process does not immediately imply that target stimuli are encoded more clearly, quickly or completely resulting in responding more quickly and accurately to targets. An increase in the bias of an encoding process for a target stimulus implies that the target need not be as clearly, quickly, or completely encoded in order to initiate a detection response. So subjects can also respond more quickly and accurately to targets by responding before having fully encoded the target. In the case of bias changes, such a facilitation of target detection is achieved at the cost of "false-alarming" to nontargets, whereas in the case of sensitivity changes it is not. A change in bias need not be a change in response bias, it can occur in either encoding as well as (postencoding) decision and response processes.

List of Literature

Allen, J. (1994) How do humans recognize speech?, *IEEE Transactions on speech and audio processing*, **2**:567-577.

Allen, J.B. (1995) *The ASA Edition of Speech and Hearing in Communication*. USA: American Institute of Physics.

Allen, J. B. (1996) Harvey Fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America*, **99**:1825-1839.

Allen, J.B., Furst, M., Saul, L. & Rahim, M. (2000) Feature extraction in human speech recognition. *Proceedings Workshop on The Nature of Speech Perception, The Psychophysics of speech perception III*, Utrecht, The Netherlands.

Altmann, G.T.M. & Steedman, M. (1988) Interaction with context during human sentence processing. *Cognition*, **30**:191-238.

American Standards Association (1960) *Acoustical Terminology*, SI. 1-1960, American Standards Association, New York.

Anderson, J.R. (1995) *Learning and memory, An integrated approach*. Singapore, John Wiley & Sons, Inc.

Andringa, T. (2001) *Continuity preserving signal processing*. PhD thesis. University of Groningen, The Netherlands.

Arai, T., Hermansky, H., Pavel, M., and Avendado, C. (1996) Intelligibility of speech with high-pass filtered time trajectories of spectral envelopes. In: *The Proceedings of the International Congres of Phonetic Sciences*. Stockholm.

Balmer, L. (1991), *Signals and Systems: An Introduction*. Prentice Hall.

Bashford, J.A., & Warren, R.M. (1987a) Multiple phonemic restorations follow the rules for auditory induction. *Perception & Psychophysics*, **42**: 114-121.

Blackburn, C.C. & Sachs, M.B. (1990) The representation of steady-state vowel sound /ε/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology*, **63**(5): 1191-1211.

Boer, E. de, (1956) *On the residue in hearing*, Unpublished doctoral dissertation, University of Amsterdam, Netherlands.

Boer, E. de, (1980) Auditory physics. Physical principles in hearing theory I. *Physics Reports*, **62**:87-174.

Boland, J.E. & Cutler, A. (1996) Interaction with autonomy: Multiple Output models and the inadequacy of the Great Divide. *Cognition* , 58:309-320.

Bourlard, H., Hermansky, H., & Morgan, N. (1996) Towards Increasing Speech Recognition Error Rates. In: *Speech Communication*, **18**: 205-231.

Bradley, L. & Bryant, P.E. (1983) Categorising sounds and learning to read -- a causal connection. *Nature* , **301**:419-421.

Bregman, A.S. & Pinker, S. (1978) Auditory streaming and the building of timbre. Canadian Journal of Psychology, **32**:19-31.

Bregman, A.S. & Rudnicky, A. (1975) Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception & Psychophysics*, **46**:395-396.

Bregman, A.S. & Steiger, H. (1980) Auditory streaming and vertical localization: Interdependence of "what" and "where" decisions in audition. *Perception & Psychophysics*, **28**:538-546.

Bregman, A.S. (1990) *Auditory Scene Analysis. The perceptual organization of sound*. Cambridge: MIT Press.

Bregman, A.S. (1998) Psychological data and computational ASA. In: D.F. Rosenthal, & H.G. Okuno (Eds.) *Computational Auditory Scene Analysis*, pp.1-12.

Broadbent, D.E. (1955) A note on binaural fusion. *Quarterly Journal of Experimental Psychology*, **7**: 46-47.

Broadbent, D.E., & Ladefoged, P. (1957) On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, **29**: 708-710.

Carhart, R., Johnson, C., & Goodman, J. (1975) Perceptual masking of spondees by combinations of talkers. *Journal of the Acoustical Society of America*, 58:S35.

Cariani, P. (1995) As if time really mattered. In: L. Rocha (Ed)., *Communication and Cognition - Artificial Intelligence (CC-AI), Special Issue on Self-Reference in Biological and Cognitive Systems*, **12**: 157-219.

Carpenter, G. A. & Grossberg, S. (1991) A massively parallel architecture for a self-organizing neural pattern recognition machine, In: G. Carpenter and S. Grossberg (Eds.), *Pattern Recognition by Self Organizing Neural Networks*, MIT Press, London, pp. 316-382.

Carlyon, R.P., Cusack, R., Foxton, J.M. & Robertson, I.H. (2000) The effects of attention and unilateral neglect on auditory streaming. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Cheesman, M.F. & Greenwood, K.G. (1995) Selective adaptation by context conditioned fricatives. *Journal of the Acoustical Society of America* , **97**:531-538.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, **13**:221-268.

Christiansen, M.H. & Chater, N. (2000) Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences.*

Clark, A. (1993) *Microcognition, Philosophy, cognitive science, and parallel distributed processing*. MIT Press.

Cleeremans, A. (2001) Learning and consciousness: A graded, dynamical perspective. In: N. Taatgen and J. Aasman (Eds.) *Third International Conference on Cognitive Modeling*. Mahwah, NJ: Erlbaum.

Connine, C.M. & Clifton, C. (1987) Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **13**: 291-299.

Connine, C.M. & Deelman, T. (1999) The psycholinguistics of spoken word recognition. In: K. Pontiag (Ed.), *Computational models of speech pattern processing* (pp.233-251). Springer-Verlag, Berlin Heidelberg.

Connine, C.M. & Titone, D. (1996). Phoneme monitoring. *Language and cognitive processes*, **11**:635-645.

Cook, P.R. (1999) *Music, Cognition, and Computerized Sound.*

Cooke, M. (1993) *Modeling Auditory Processing and Organization*. Cambridge: Cambridge University Press.

Costa, A., Cutler, A. & Sebastian-Galles, N. (1998). Effects of phoneme repertoire on phoneme decision. *Perception & Psychophysics*, **60**:1022-1031.

Crowder, R.G. (1981) The role of auditory memory in speech perception and discrimination. In. T. Myers, J. Laver, & J. Anderson (Eds.) *The cognitive representation of speech* (pp.167-179). New York: North-Holland.

Cutler, A. & Otake, T. (1994) Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, **33**:824-844.

Cutting, J.E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**:114-140.

Dai, H. & Wright, B.A. (1995) Detecting signals of unexpected or uncertain durations. *Journal of the Acoustical Society of America*, **90**:1371-1380.

Dai, H., Neff, D.L & Jesteadt, W. (2000) Psychometric functions for detecting level increments added to continuous or gated pedestals. *Journal of the Acoustical Society of America.*

Dai, H., Neff, D.L., Jesteadt, W. & Neely, S. (2000) An integrated look at auditory temporal integration models. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Dalenoort, G., & de Vries, P. (1995a) *Algemene Systeem Theorie*. Unpublished manuscipt. Rijksuniversiteit Groningen.

Dalenoort, G., & de Vries, P. (1995b) *Cognitieve Systemen en Modellen*. Unpublished manuscipt. Rijksuniversiteit Groningen.

Dallos, P. (1988) Cochlear neurobiology: Some key experiments and concepts over the past two decades. In: G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.) *Auditory function, Neurobiological bases of hearing*, pp. 153-187.

Dannenbring, G.L. (1976) Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, **30**:90-114.

Dannenbring, G.L. & Bregman, A.S. (1978) Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics*, **24**:369-376.

Darwin, C.J. (1981) Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33A**:185-208.

Darwin, C.J. & Carlyon, R.P. (1995) Auditory grouping. In: B.C.J. Moore (Ed.) *Hearing, Handbook of Perception and Cognition*, pp.387-424. San Diego: Academic Press.

Darwin, C.J., Howell, P., & Brady, S.A. (1978) Laterality and localization: A right ear advantage for speech heard on the left. In:. J. Requin (Ed.), *Attention and Performance VII*. Hillsdale, N.J.: Erlbaum.

Darwin, C.J. & Sutherland, N.S. (1984) Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology*, 36: 193-208.

Davis, M.H. (2000) *Lexical segmentation in spoken word recognition*. PhD thesis. Birkbeck college, University of London.

Davis, M. H., Marslen-Wilson, W. D. & Gaskell, M. G. (1997) Ambiguity and competition in lexical segmentation. In: M.G. Shafto and P.Langley (Eds.) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

de Cheveigné, A. (1998) Cancellation model of pitch perception. *Journal of the Acoustical Society of America*, **104**: 1118-1121.

de Cheveigné, A. (2000) The Auditory System as a Separation Machine. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Delgutte, B. (1997) Auditory neural processing of speech. In: W.J. Hardcastle, & J. Laver (Eds.) *Handbook of Phonetic Science*. Blackwell, Oxford.

Delgutte, B. & Kiang, N.Y.S. (1984) Speech encoding in the auditory nerve. I. Vowel-like sounds. *Journal of the Acoustical Society of America*, 75: 866-878.

Delgutte, B. & Kiang, N.Y.S. (1984b) Speech coding in the auditory nerve. IV. Sounds with consonantlike dynamic characteristics, *Journal of the Acoustical Society of America*, 75:897-907.

Drullman, R., Festen, J.M., and Plomp, R. (1994) Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95: 1053-1064.

Duifhuis, H. (1973) Consequences of peripheral frequency selectivity for nonsimultaneous masking. *Journal of the Acoustical Society of America*, **54**:1471-1488.

Duifhuis, H. (1976) Cochlear nonlinearity and second filter: possible mechanism and implications. *Journal of the Acoustical Society of America*, **59**:408-423.

Eimas, P.D., Marcovitz Hornstein, S.B. & Payton, P. (1990) Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language* , **29**:160-180.

Eimas, P.D. & Nygaard, L.C. (1992) Contextual coherence and attention in phoneme monitoring. *Journal of Memory and Language* , **31**:375-395.

Eggermont, J.J. (2000) Across-perceptual channel gap detection properties revealed in single auditory cortical neurons, *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing* (pp. 313-319). Eindhoven, The Netherlands.

Ellis, D.P.W. (1996) *Prediction-driven Computation Auditory Scene Analysis*. PhD Thesis. Cambridge, MIT Press.

Ellis, D.P.W. (1996) Prediction-driven computational Auditory Scene Analysis for dense sound mixtures. *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception*. Keele, UK.

Elman, J.L. & McClelland, J.L. (1988) Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language* , **27**:143-165.

Farah, M. (1989) Semantic and perceptual priming: How similar are the underlying mechanisms? *Journal of Experimental Psychology: Human Perception and Performance* , **15**:188-194.

Fletcher, H. (1940) Auditory patterns. *Review Modern Physics*, **12**:47-65.

Fodor, J.A. (1983) *The modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. & Bever, T.G. (1965) The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behaviour*, **4**:414-420.

Forster, K.I. (1979) Levels of processing and the structure of the language processor. In: W.E. Cooper & E.C.T. Walker (Eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 27-85). Erlbaum.

Forster, K.I. (1981) Priming and the effects of sentence and lexical context on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology A*, **33**: 465-495.

Foss, D.J. (1982) A discourse on semantic priming. *Cognitive Psychology*, **14**: 590-607.

Foss, D.J. & Blank, M.A. (1980) Identifying the speech codes. *Cognitive Psychology* , **12**:1-31.

Frazier, L. (1987) Theories of syntactic processing. In: J.L. Garfield (Ed.) *Modularity in knowledge representation and natural language processing* (pp. 291-307). Cambridge, MA: MIT Press.

Ganong, W.F. (1980) Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* , **6**:110-125.

Gold, B. & Morgan, M. (2000) *Speech and audio signal processing. Processing and Perception of Speech and Music.* .....

Goldinger, S.D., Luce, P.A., Pisoni, D.B. & Marcario, J.K. (1992) Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition* , **18**:1211-1238.

Goldstein, E.B. (1989) *Sensation and Perception*. Pacific Grove, California, Brooks/Cole Publishing Company.

Goldstein, J.L. (1973) An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, **54**:1496-1516.

Green, D.M. (1976) *An introduction to Hearing*. Hillsdale, New Jersey.

Greenberg, S. (1988) The ear as a speech analyser. *Journal of Phonetics*, 16:139-150.

Greenberg, S. (1995) The ears have it: The auditory basis of speech perception. In: *The Proceedings of the International Congres of Phonetic Sciences* (pp. 34-41). Stockholm.

Greenberg, S. (1996) Understanding speech understanding – Towards a unified theory of speech perception. *Proceedings of the ESCA Tutorial of Advanced Research Workshop on the Auditory Basis of Speech Perception*. Keele, England, pp. 1-8.

Grossberg, S., (1991) Temporal order and word recognition. In: G. Carpenter and S. Grossberg (Eds.), *Pattern Recognition by Self Organizing Neural Networks*, MIT Press, London, pp. 593-610.

Haggard, M.P, Hall, J.W. & Fernandes, M.A. (1984) Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76**:50-56.

Hall, J.W., Haggard, M.P. & Harvey, A.D.G. (1984) Release from masking through ipsilateral and contralateral comodulation of a flanking band. *Journal of the Acoustical Society of America*, **76**:S76.

Halligan, P.W. & Marshall, J.C. (1993) The history and clinical presentation of neglect. In: I.H. Robertson, & J.C. Marshall (Ed.). *Unilateral Neglect: Clinical and Experimental Studies*. Lawrence Erlbaum, Hove (pp.3-26).

Hartmann, W.M. (1988) Pitch perception and the segregation and integration of auditory entities. In: G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.) *Auditory Function, Neurobiological Bases of Hearing*, (pp.623-646).

Helmholtz, H.L.F. (1885) *On the Sensations of Tone* (English Edition), Longmans.

Houkema, D. (1999a) *Een recensie over het boek: Mind as Motion: An exploration in the dynamics of cognition*. Unpublished report. Rijksuniversiteit Groningen.

Houkema, D. (1999b) *Spraakherkenning*. Unpublished report. Rijksuniversiteit Groningen.

Houkema, D., & Warris, S. (1999) *Parallel Distributed Processing, Practicumverslag Neuroinformatica*. Unpublished report. Rijksuniversiteit Groningen.

Huckvale, M.A. (1996) Learning from the experience of building automatic speech recognition systems. In: *Speech Hearing and Language - Work in Progress, Phonetics and Linguistics*, University College London.

Huckvale, M.A. (1997) 10 Things engineers have discovered about speech recognition. *Proceedings of the NATO ASI Speech Pattern Processing*, Jersey, 1997.

Huckvale, M.A (1998) Opportunities for re-convergence of engineering and cognitive science accounts of spoken word recognition. *Proceedings of the Institute of Acoustics*, **20**: 9-20.

Huggins, A.W. (1975) Temporally segmented speech. *Perception & Psychophysics*, 18:149-157.

Ifeachor, E.C. & Jerwvis, B.W. (1993), *Digital Signal Processing: A Practical Approach*. Addison-Wesley.

Johnstone, B.M., Patuzzi, R. & Yates, G.K. (1986) *Hearing Research*, **22**: 147–153.

Julesz, B. & Hirsch, I.J. (1972) Visual and auditory perception: An essay of comparison. In: E.E. David, & P.B. Denes (Eds.) *Human communication: A unified view* (pp.283-340). New York: McGraw-Hill.

Kess, J.F. (1992), *Psycholinguistics. Psychology, linguistics and the study of natural language*. Amsterdam, John Benjamins Publishing Co.

Kim, J. (1996) Philosophy of Mind. Oxford: Westview Press.

Kingsbury, B.E.D., Morgan, N., and Greenberg, S. (1988) Robust speech recognition using the modulation spectrogram, *Speech Communication*, **25**(1-3): 117-132.

Klatt, D.H. (1989) Review of selected models of speech perception. In: W. Marslen-Wilson (ed.) *Lexical Representation and Process* (pp. 169-226). Cambridge: MIT Press.

Kollmeier, B. & Derleth, R.P. How linear is the auditory system? Models of compression and expansion based on psychoacoustics for normal and hearing-impaired listeners. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Kubovy, M. & Jordan, R. (1979) Tone segregation by phase: On the phase sensitivity of the single ear. *Journal of the Acoustical Society of America*, **66**:100-106.

Ladefoged, P. & Broadbent, D.E. (1960) Perception of sequence in auditory events. *Quarterly Journal of Experimental Psychology*, **12**:162-170.

Langner, G. & Schreiner, C. (1988) Periodicity coding in the inferior colliculus of the cat. *Journal of Neurophysiology*, **60**: 1805-1822.

Lehiste, I. (1970) *Suprasegmentals*. Cambridge: MIT Press.

Liberman, A.M. (1982) On finding that speech is special. *American Psychologist*, **37**:148-167.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review*, *74*:431-461.

Liberman, A.M. & Mattingly, I.G. (1989) A specialization for speech perception. *Science*, **243**:489-494.

Liberman, A.M. & Studdert-Kennedy (1978) Phonetic perception. In: R. Held, H. Leibowitz, & H.-L. Teuber (Eds.) *Handbook of sensory physiology: Vol. 8 Perception* (pp.143-178). New York, Springer-Verlag.

Lyon, R.F. (1984) Computational models of neural auditory processing. *IEEE ICASSP*, 3611-3614.

Mammano, F. & Nobili, R. (1993) *Journal of the Acoustical Society of America*, **93**: 3320–3332.

Mann, V.A. & Repp, B.H. (1981) Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, **69**:548-558.

Marin, C. (1987) Rôle d'enveloppe spectrale dans la perception des sources sonores. *Memoire de la DEA de Phonétique*, Université de la Sorbonne Nouvelle, Paris III.

Marslen-Wilson, W. D., Ford, M., Older, L., & Zhou, X. (1996). The combinatorial lexicon: Affixes as processing structures. In G. W. Cottrell (Ed). *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Marslen-Wilson, W. & Tyler, L. (1980) The temporal structure of spoken language understanding. *Cognition*, 8: 1-71.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, **101**: 3-33.

Marslen-Wilson, W. & Warren, P. (1994) Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review* , **101**: 653-675.

Massaro, D.W. (1987) *Speech perception by ear and eye: A paradigm for psychological enquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D.W. (1989a) Testing between the TRACE model and the Fuzzy Logical Model of Speech Perception. *Cognitive Psychology*, **21**:398-421.

Massaro, D.W. (1989b) Multiple book review of "Speech perception by ear and eye: A paradigm for psychological enquiry". *Behavioral and Brain Sciences*, **12**:741-794.

Massaro, D.W. (1996) Integration of multiple sources of information in language processing. In: *Attention and performance XVI: Information integration in perception and communication*, ed. T. Inui & J.L. McClelland (pp. 397-432). MIT Press.

Massaro, D.W. (1997). *Perceiving talking faces: From speech perception to a behavioral principle* . MIT Press.

Massaro, D.W. & Cohen, M.M. (1991) Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology* , **23**:558-614.

Massaro, D.W. & Oden, G.C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition* , **21**:1053-1064.

Mattingly, I.G. & Liberman, A.M. (1990) Speech and other auditory modules. In: G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.) *Signal and Sense: Local and global order in perceptual maps* (pp.501-520). New York: Wiley.

McAdams, S. (1984) *Spectral fusion, spectral parsing, and the formation of auditory images*. Doctoral dissertation, Stanford University.

McAdams, S. (1993) Recognition of sound sources and events. In: S. McAdams & E. Bigand (Eds.) *Sound: The Cognitive Psychology of Human Audition* (pp.146-198). Oxford: Oxford University Press.

McAdams, S. & Rodet, X. (1988). The role of FM-induced AM in dynamic spectral profile analysis. In: H. Duifhuis, J. Horst, and H. Wit (Eds.), *Basic Issues in Hearing*. London: Academic Press.

McClelland, J.L. & Elman, J.L. (1986) The TRACE model of speech perception. *Cognitive Psychology*, **18**:1-86.

McClelland, J.L. (1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* , **23**:1-44.

McClelland, J.L., St. John, M. & Taraban, R. (1989) Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4:287-336.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* (London), **264**:746-748.

McQueen, J. (1996) Phonetic categorisation. *Language and Cognitive Processes* , **11**:655-664.

McQueen, J.M. & Cutler, A. (1997). Cognitive processes in spoken-word recognition. In W.J. Hardcastle & J.D.M.H. Laver (Eds.) *A Handbook of Phonetic Science*. Oxford: Blackwell, 566-585.

McQueen, J.M., Cutler, A., Briscoe, T. & Norris, D. (1995) Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes* , **10**:309-331.

McQueen, J.M., Cutler, A. and Norris, D. (1999) Lexical activation produces impotent phonemic percepts. *138th Meeting of the Acoustical Society of America*. Columbus, OH, November.

McQueen, J.M., Norris, D.G. & Cutler, Al. (1999) Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, 25:1363-1389.

Meddis, R. & Hewitt, J. (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, **89**: 2866-2882.

Miller, G.A., & Licklider, J.C. (1950) The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, 22:167-173.

Moore, B.C.J. (1977) *The Psychology of Hearing*. Cambridge: University Press.

Morais, J., Cary, L., Alegria, J. & Bertelson, P. (1979) Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* , 7:323-331.

Morgan, D.P. & Scofield, C.L. (1991) The mammalian auditory system. In: *Neural Networks and Speech Processing*, Ch2 (pp.9-40).

Morton, J. (1969) Interaction of information in word recognition. *Psychological Review*, **76**: 165-178.

Neely, S.T. & Dai, H. (2000) Inferring properties of IHC excitation from psychophysical growth of masking. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Nobili, R., Mammano, F. & Ashmore, J. (1998) How well do we understand the cochlea? *Trends in Neurosciences*, **21**:159-167.

Nooteboom, S.G. & Cohen, J.A. (1995) *Spreken en verstaan*.

Norris, D. (1982). Autonomous processes in comprehension: A reply to Marlsen-Wilson and Tyler. *Cognition*, **11**:97-101.

Norris, D. (1993) Bottom-up connectionist models of "interaction". In: G.T.M. Altmann & R. Shillcock (Eds.) *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 211-234). Erlbaum.

Norris, D.G. (1994b) Shortlist: A connectionist model of continuous speech recognition. *Cognition* , **52**:189-234.

Norris, D. (1995) Signal detection theory and modularity: On being sensitive to the power of bias models of semantic priming. *Journal of Experimental Psychology: Human Perception and Performanc* , 2:935-939.

Norris, D., McQueen, J.M. & Cutler, A. (1995) Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* , **21**:1209-1228.

Norris, D., McQueen, J.M., Cutler, A. & Butterfield, S. (1997) The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34:191-243.

Norris, D., McQueen, J.M. & Cutler, A. (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23.

O'Grady, W., Dobrovolsky, M. & Aronoff, M. (1991) *Contemporary linguistics, An introduction*. New York: St. Martin's Press.

Peeters, G., Frauenfelder, U. & Wittenburg, P. (1989) Psychological constraints upon connectionist models of word recognition: Exploring TRACE and alternatives. In: R. Pfeifer, Z. Schreter, F. Fogelman-Soulié & L. Steels (Ed.) *Connectionism in perspective* (pp. 395-402). Elsevier.

Pisoni, D.B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**:253-260.

Plack, C.J. & White, L.J. (2000) Temporal integration in pitch perception. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Plack, C.J. & White, L.J. (2000) Perceived continuity and pitch perception. *Journal of the Acoustical Society of America* (in press).

Plaut, D.C. (2000) Methodologies for the computer modelling of human cognitive processes. In: F. Boller and J. Grafman (Eds.) *Handbook of Neuropsychology*, 2nd Edition. New York: Elsevier.

Pols, L.C.W. (1997b) Flexible, robust, and efficient human speech recognition. *IFA Proceedings*, **21**: 1-10.

Port, R.F. & Gelder, T. van (1995) *Mind as Motion, Explorations in the dynamics of Cognition*, Cambridge, MA: MIT-press.

Remez, R.E., Rubin, P.E. (1983) The stream of speech. *Scandinavian Journal of Psychology*, **24**:63-66.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981) Speech perception without traditional speech cues. *Science*, **212**:947-950.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S. & Lang, J.M. (1994) On the perceptual organization of speech, *Psychological Review*, **101**: 129-156.

Repp, B.H. (1980) Accessing phonetic information during perceptual integration of temporally distributed cues. *Journal of Phonetics*, **8**:185-194.

Robles, L., Ruggero, M.A. & Rich, N.C. (1986) *Journal of the Acoustical Society of America*, **80**: 1364–1374.

Rose, J.E., Brugge, J.F., Anderson, D.J. & Hind, J.E. (1967) Phase-locked response to low-frequency tones in single auditory nerve fibres of the squirrel monkey. *Journal of Neurophysiology*, **30**:769-793.

Rubin, P.E., Baer, T. & Mermelstein, P. (1981) An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, **70**:321-328.

Sachs, M.B. (1984) Speech encoding in the auditory nerve. In: C.I. Berlin (Ed.), *Hearing Science Recent Advances*. College-Hill, San Diego, CA (pp.263-307).

Sachs, M.B., Blackburn, C.C., & Young, E.D. (1988a) Rate-place and temporal-place representation of vowels in the auditory nerve and anteroventral cochlear nucleus. *Journal of Phonetics*, **16**:37-53.

Sachs, M.B., Winslow, R.L., & Blackburn, C.C. (1988b) Representation of speech in the auditory periphery. In: G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.) *Auditory Function, Neurobiological Bases of Hearing*, pp.747-774.

Saltzman, E.L. (1995) Dynamics and coordinate systems in skilled sensorymotor activity. In: R.F. Port, & T. van Gelder (Eds.) *Mind as Motion, Explorations in the dynamics of Cognition*, Cambridge, MA: MIT-press, Ch6 (pp.149-173).

Samuel, A. (1996) Phoneme restoration. *Language and Cognitive Processes* , **11**:647-654.

Samuel, A.G. (1996) Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General*, **125**:28-51.

Samuel, A.G. & Kat, D. (1996) Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance* , **22**:676-694.

Schouten, J.F. (1940) *Five articles on the perception of sound (1938-1940)*, Institute for Perception, Eindhoven, Netherlands.

Schouten, J.F. (1970) The residue revisited. In: R. Plomp, & G.F. Smoorenburg, *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, Leiden.

Schreiner, C. and Urbas, J.V. (1988) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, **21**: 227-241.

Scott, B. & Cole, R. (1972) Auditory illusions as a function of imbedded sounds. *Journal of the Acoustical Society of America*, **51**:112 (Abstract).

Segui, J., Dupoux, E., Mehler, J. (1990) The role of the syllable in speech segmentation, phoneme identification, and lexical access. In: G. Altmann (Ed.) *Cognitive Models of Speech Processing: Psycholinguistic and computational perspectives* (pp.263-280). Cambridge: MIT Press.

Shamma, S., Klein, D. & Depireux, D. (2000) Coincidence detection in pitch perception. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Shannon, R.V. Zeng, F-G., Kamath, V., and Wygonski, J. (1995) Speech recognition with primarily temporal cues. Science, **270**: 303-304.

Skoyles, J.R. (1998) Speech phones as a replication code. In: *Medical Hypotheses*, **50**: 167-173.

Slaney, M. & Lyon, R. (1993) On the importance of time – A temporal representation of sound. In: M. Cooke, S. Beet and M. Crawford (Eds.) *Visual Representations of Speech Signals*. J. Wiley and Sons, Sussex England.

Slaney, M. (1998) A critique of pure audition. In: Rosenthal, D.F. & Okuno, H.G. (Eds.) Computational Auditory Scene Analysis (pp.27-42).

Smith, R.L., Baptiste, D. & Lütkenhöner, B. (2000) Incremental responses in the auditory periphery and the auditory cortex, and a possible relationship to psychophysical overshoot. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Smoorenburg, G.F. (1992) Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America*, **91**:421-437.

Steiger, H. (1980) *Some informal observations concerning the perceptual organization of patterns containing frequency glides*. Unpublished report. McGill University.

Suga, N. (1988) Auditory neuroethology and speech processing: Complex-sound processing by combination-sensitive neurons. In: G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.) *Auditory Function, Neurobiological Bases of Hearing*, pp.679-720.

Tchorch, J. & Kollmeier, B. (1999) A model of auditory perception as front end for automatic speech recognition. *Journal of the Acoustical Society of America*, **106**: 2040-2050.

Terhardt, E. (1970) Frequency analysis and periodicity detection in the sensations of roughness and periodicity pitch. In: R. Plomp, & G.F. Smoorenburg, *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, Leiden.

Terhardt, E. (1974) Pitch, consonance and harmony. *Journal of the Acoustical Society of America*, **55**:1061-1069.

Tougas, Y. & Bregman, A.S. (1985a) Auditory streaming and the continuity illusion. *Perception & Psychophysics*.

Traunmüller, H. (2000) The modulated nature of speech: Implications for the interpretation of some experimental results. *Proceedings Workshop on The Nature of Speech Perception, The Psychophysics of speech perception III*, Utrecht, The Netherlands.

Treisman, A. (1987) Properties, parts, and objects. In: K.R. Boff, & L. Kaufman, & J.P. Thomas (Eds.) *Handbook of perception and human performance* (Ch35). New York: Wiley.

Van Noorden, L.P.A.S. (1975) *Temporal Coherence in the Perception of Tone Sequences.* Unpublished doctoral dissertation, Eindhoven University of Technology.

Viemeister, N.F. & Wakefield, G.H. (1991) Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, **90**:858-865.

Vitevitch, M.S. & Luce, P.A. (1998) When words compete: Levels of processing in spoken word recognition. *Psychological Science* , 9:325-329.

von Békésy, G. (1960) *Experiments in Hearing*, McGraw–Hill.

Wang, X., Lu, T. & Liang, L. (2000) Temporal processing in the primate auditory cortex. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Warren, R. M., & Sherman, G. L. (1974) Phonemic restorations based on subsequent context. *Perception and Psychophysics*, **16**(1): 150-156.

Warren, R.M. (1982) *Auditory Perception: A New Synthesis.* New York: Pergamon.

Weintraub, M. (1985) *A Theory and Computational Model of Monaural Auditory Sound Segregation*, PhD Thesis, Stanford University.

White, L.J. & Plack, C.J. (1998) Temporal processing of the pitch of complex tones. *Journal of the Acoustical Society of America*, **103**: 2051-2063.

Whitfield, I.C. (1970) Central nervous processing in relation to spatio-temporal discrimination of auditory patterns. In: Plomp, R. & Smoorenburg, G.F. *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, Leiden.

Wightman, F.L. (1973b) The pattern transformation model of pitch. *Journal of the Acoustical Society of America*, **54**: 417-426.

Wurm, L.H. & Samuel, A.G. (1997) Lexical inhibition and attentional allocation during speech perception: Evidence from phoneme monitoring. *Journal of Memory and Language*, **36**:165-187.

Young, S & Knill, K. (1997) Hidden Markov Models. In: S. Young, & G. Bloothooft (Eds.) *Corpus-based Methods in Speech and Language Processing*, Ch2 (pp.27-68).

Yu, J.J. & Young, E.D. (2000) Rate-place representation and feature detection: Two mechanisms of spectral information transmission by cochlear nucleus neurons. *Physiological and Psychophysical Bases of Auditory Function, Proceedings of the 12th International Symposium on Hearing*. Eindhoven, The Netherlands.

Zwicker, E. & Scharf, B. (1965) A model of loudness summation. *Psychological Review*, **72**:3-26.

Zwicker, E. & Fastl, H. (1999) *Psychoacoustics. Facts and models.* Berlin Heidelberg: Springer-Verlag