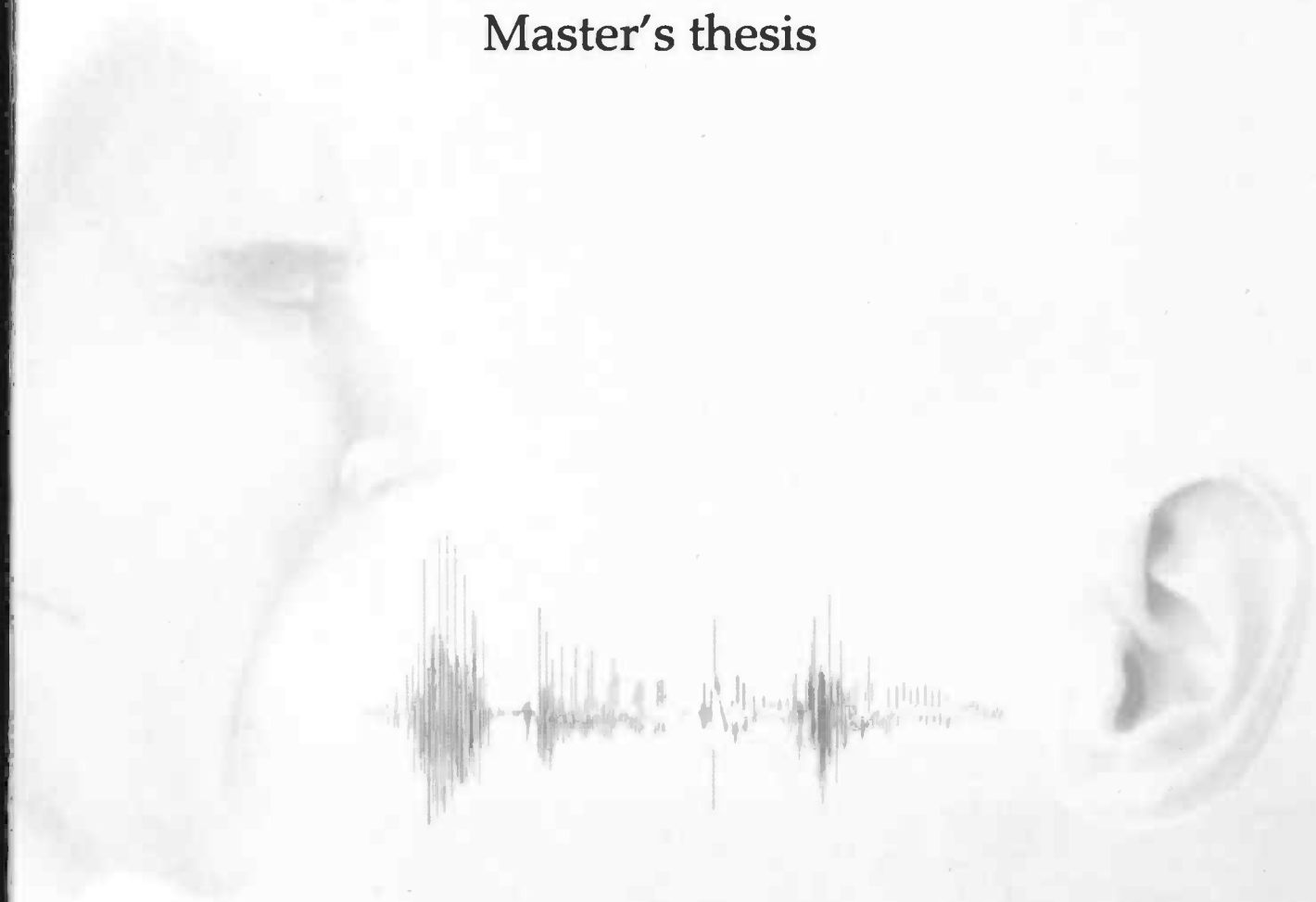
On the Synthesis
of
Aggressive Vowels

Towards more robust aggression detection.

Master's thesis



Joep Boers
August 2006

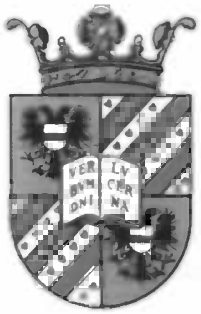
957

On the Synthesis of Aggressive Vowels:
Towards more robust aggression detection

Joep Boers
s1288873

Supervisors:

- Dr. T.C. Andringa, *RuG / KI*
- Drs. M. Huisman, *Sound Intelligence*



Sound Intelligence
Sint Jansstraat 2
9712 JN Groningen

Rijks*universiteit* Groningen
Faculteit der Gedrags- en Maatschappijwetenschappen
Afdeling Kunstmatige Intelligentie
Grote Kruisstraat 2/1
9712 TS Groningen

957.

Contents

List of Figures

Acknowledgments

Let every man judge according to his own standards, by what he has himself read, not by what others tell him.

Albert Einstein 1879–1955

Hear, hear! I have made it, I tell you! However, it feels like being in the "Weekend miljonairs" quiz. Finishing my master is like having answered a bunch of questions correctly, and reaching a level at which you are certain you will go home with a great price. But there are still more questions to come. There is always a next level and the questions will be harder to answer. But, answering them will bring you closer to the million dollar reward ...

The last period of my study was rather tough: one of the drawbacks of not being an average student, in the sense that I already finished highschool like a few years back, is, certainly, a more complicated social life. If any. However, being the kind who likes to win a regatta in his ancient boat, before accepting new, though common rigging, I am aware of some paradoxes of life. Still, I wouldn't want it the other way around.

I owe my gratitude to a few people. First and most of all, I would like to thank my girlfriend, and soul mate, Bianca, for her patience and inspiration the last five years. Besides kicking my ass when needed, she would be the one who understands my innermost motivations. I also like to thank my supervisors, Dr. Tjeerd Andringa and Drs. Mark Huisman, for their guidance during my graduation project. Further, I very much appreciate the fact that Dr. ir. Peter van Hengel sympathized with my private struggles. During the past six months I had the opportunity to discuss matters with a few other experienced scientists. I discovered that it can be very fruitful to ask these people questions. I tend to do it all on my own, which, in a sense, might be a very dangerous attitude. Brainstorming with Dr. Esther Wiersinga-Post, Prof. Dr. Veldman, and Prof. Dr. Schutte turned out to be very enlightening. Finally, without mentioning their names, I would very much like to show my respect to two dear friends, stemming from the days I led a happy life sailing on the "Bruine Vloot". These smart people were real life-motivators.

Joep Boers

Vierhuizen, July 2006

Abstract

Sound and speech recognition are important research areas in artificial intelligence. Humans are very well able to detect aggression in verbal expressions. Knowledge of the relation between emotions, e.g. aggression, and acoustic features in speech may be of much use improving, for instance, speech recognition. Currently, Sound Intelligence is working on the development of the next generation of aggression detectors. Those systems are aimed at not only detecting aggression, but also classifying verbal expressions of human aggression(in real-life circumstances).

Much research is done on the perceptual side of the speech chain. However, in order to come to aggression classification we focus on the speech production of Dutch vowels. Parallel to human speech production, we developed, implemented, and evaluated a vocoder which was used to synthesize vowels intended to exhibit gradations of emotions, primarily aggression. In contrast to former research on human recognition of verbal emotions, normally conducted on genuine, rich and labeled data we defined cues and subsequently synthesized vowels. By means of a psycho-acoustic experiment we believe to have proven that this approach, and thus the vocoder, is scientific justified. Still, aggression classification needs much more further research –and it is our belief that nonlinear analysis might be very useful here (literature shows very interesting progress)– but either way, a vocoder, like the one used in this work, is expected to be complementary to current research approaches.

Chapter 1

Introduction

No, I am not angry about anything –I just cry all the time.

John Doe

Humans are very well able to detect aggression in verbal expressions. We normally do not need to observe someone's facial expressions to come to the conclusion that he or she is in a very aroused state of mind. Nor do we have difficulties in detecting the change of his or her arousal. When you are teasing your friend, you know when to stop. Her voice gives you clues as to when her meek swallowing turns into an 'enough is enough' situation. Her voice clearly changes pitch and when you stubbornly keep on teasing her, her voice may change into a trembling kind of lion-like roaring. At that stage you know you have gone to far: why didn't you stop annoying her when she gave you her clear warnings?

This sketches the ease with which we are able to analyze vocal expressions. This gift we all are aware off, will normally help us to act appropriate in many given situations. Of course, in practice we will combine evidence from multiple sources, that is, use facial expressions too, but to a wide extent this often is not necessary. Now, there are many occasions in which it would be very helpful to detect upcoming anger automatically. We, then, would like to intervene before anger becomes aggression. For this task it is possible to build an aggression detector. However, an aggression detector is a kind of binary device: it tells you when it came to the conclusion that there is aggression. What we really would like our detector to do is to *classify* aggression. We would like to have some measure of the amount of aggression present in verbal sound. Then we are able to react appropriately in a given situation and prevent that situation to escalate. Unfortunately, classification turns out to play hard to get.

The question now is what in a voice makes us aware of the different levels of an emotion? What happens to a verbal utterance when one becomes more and more aroused? To answer this question we have to take a look at the production of speech. Of course, the production of a complete sentence is a very complex matter. But becoming angry is, in a sense, losing grip on the carefully considered manufacturing of speech. Assuming this, one would expect that primarily physiological changes would affect our production of speech in the case of developing aggression. On the other hand, before one has completely lost control of her verbal finesse, one will probably use some intonation of voice, consciously, to raise the warning flags.

Literature has come up with quite a few parameters of vocal properties related to the diverse states of emotions. Still, these have not been convenient enough to

come to a decent classification, when at all possible. Furthermore, most literature has long been ignoring the role of the sound producing organ. It seems that it's role has been underestimated, or, at least, it's role is likely to be of great importance in the task of detecting or classifying emotions like aggression. In this work we examine some phenomena resulting from the (nonlinear) behavior of the vocal fold. These phenomena, especially the shape of the pulse train produced by the vocal folds, and irregularities like jitter and shimmer, will be evaluated on their utility. Jitter and shimmer are, in this text, defined as frequency and amplitude variations, theoretically induced by increasing velocity of the airflow which is the power source for the realization of vowels and consonants. To gain knowledge about the importance of these effects a vocoder is build. This vocoder can be adjusted such that it produces vowels with, hopefully, an aggressive content. It can be used to test the amount of aggressiveness perceived by test subjects. Instead of relying on recordered and subsequently labeled speech fragments, followed by analysis and, in a way, reducing it's richness of spectral content, one could test certain hypotheses and gradually build up sound until it approaches human quality.

This work, partly conducted at SOUND INTELLIGENCE, is organized in the following manner. Chapter 2 gives a short overview of research on the subject of observing emotions in speech, like aggression. Some ideas are unfolded on what to look at. Moreover, the physiology of the voice producing organ will be discussed in more detail, taking into account recent knowledge. In this chapter a few words are spent on known acoustic cues. Here also the Teager Enery Operator is introduced. We expect to be able to investigate acoustical cues with it, stemming from the nonlinear behavior of the sound producing system. Chapter 3 is about the research objectives. In chapter 4 the source filter model and the vocoder are discussed. Modeling of the glottis is regarded in a chapter of it's own, because of it's importance: chapter 5. The last two chapters explain and discuss the conducted experiment (chapter 6) and subsequently evaluate the results obtained. Also a glimpse of our thoughts on future work is put into written words (chapter 7).

Chapter 2

Theoretical background

It would be a considerable invention indeed, that of a machine able to mimic speech, with its sounds and articulations. I think it is not impossible.

Leonhard Euler (1761)

In order to come to a working model for classification of aggressive vowels we first have to consider the speech chain. Speakers produce sounds and transmit them via their lips through the air as a medium. Listeners then, hopefully, hear and understand the verbal utterance of the speaker. In the speech chain one recognizes *speech production* on the one side and *speech perception* on the other side. Included in a complete speech chain are also the intentions of the speaker, for which she tries to find the words to utter in a way the listener will understand (i.e. language), and the processing of air-pressure disturbances to recognizing structure in it and understanding the message. Transmissions through a medium, connecting speakers and hearers and playing a decisive role in the speech chain, are subject to phenomena as noise, reverberation, interferences, et cetera. It is evident that there are many aspects of interest and that they can all be of significant importance. In order to be able to focus on those aspects of consequence for this work, one has to consider what ones goals are. As mentioned before, ultimately we want to come to a classification of aggression. Being able to synthesize aggressive vowels is a means to reach that goal, since we then could carry out experiments in which test subjects are asked to judge the aggressive content of an utterance, while we already know what spectral fingerprint is present. The results of such experiments would enable us to optimize our model(s), and extracting parameters from it may allow us to improve our aggression detection software. To accomplish the latter, we could try to think of new acoustical parameters and subsequently process a database of expressions of aggression and, for instance, apply statistical methods on the results. Depending on the 'quality' of the database we might stumble on decisive parameters and conclude our work ended to be successful. Unfortunately such an approach does not help us in understanding the mechanisms of the influence of emotion on speech, per se. Therefore we will first take a closer look at different aspects of the speech chain, in order to gain a better grasp on the mechanisms of most importance, but the focus will be kept on the speech production.

Speech production is thought of as independent of perception. Notwithstanding that evolution may have been the architect of our speech producing organs as well

as our hearing ability, and it may also have orchestrated human verbal communication by tuning both non-independently as to achieve the robust performance it exhibits today, perception is expected to result in understanding the message. A speaker then could alter the content of her message or intensify it, depending on the *anticipated* reaction of the listener. This means that in establishing the aggressive content of some message, we can ask a listener to put into words her perception of the utterance. However, this consideration brings us to the need to be able to *describe* aggression, or any emotion in general, to be able to compare results.

We will proceed by discussing some more theory involving emotion, like aggression, in section 2.1. The production of speech will be discussed in more detail in section 2.2. Acoustic cues known from literature will be summarized in chapter 2.3. Possible methods of analysis will be reviewed in section 2.4, where also some models of the speech producing system are mentioned. Can we use the predictions by numerical models as a guidance to search for clues of aggression in human speech? We expect they will. Finally, an expected fruitful method for analysis will be introduced in section 2.5. Researchers are currently extending and optimizing this method in the context of emotion recognition.

2.1 About emotion

A spoken message carries more information than its written counterpart. Speech gives us information about the gender and age of the speaker, as well as her regional background, intentions, attitudes, and emotional and health state [37]. Also the situation and topic of conversation leave behind their fingerprints on speech. The first concern for a speech signal is of course that it contains the message that someone wants to send. Besides that, it is structured in such a way that a listener can extract other information. By means of varying ones intonation one can emphasize parts of a sentence as to stress the importance of it. It also structures the phrasing of a sentence or dialogue (ibid). This information, not included in the syntactical or lexical content of the words, is assigned by means of *prosody*. Section 2.3.1 will pay a little more attention on the subject of prosody.

Modeling variability, i.e. the effects mentioned above, involves understanding how speech variations are performed and perceived [37]. Our goal is to use this understanding to improve aggression detection (or perhaps to make automatic speech recognition more robust). When emotions rise high, it might be that the intentional structuring, probably unconsciously, gets partly obscured by more chaotic, uncontrolled sound. When someone displays hot anger, her voice may start to tremble, the rate of speech may increase, and the lungs may pump their value content wildly through the vocal tract, making the airflow, which normally would flatter our ears with pleasant and harmonic vowels, punish our vocal folds in furious oscillations. Before this uncontrollable use of voice will be practiced, a speaker, in general, will

utilize her instrument to signal her increased emotional state. She might succeed in doing so by, for instance, raising pitch such that pitch frequency is close to the first formant frequency. This will result in more energy emitted with less effort [50]. (Not changing pitch but just trying to make speech louder would be less profitable.)

In literature there is not a widely accepted definition of emotion. An early review of definitions was given by Plutchik (1980). He discusses several theories of emotion, starting with the ideas of four pioneers: Charles Darwin, William James, Walter Cannon, and Sigmund Freud. They were mostly concerned with evolutionary, psychophysiological, neurological and dynamic approaches, respectively. That is, the evolutionary benefit of emotional behavior, the relation with bodily changes, the relation with brain structures and processes, and the meanings of unconscious and mixed emotions of people. Plutchik continues with more recent ideas about the nature of emotions and concludes with the proposal of a definition of emotion:

> *An emotion is an inferred complex sequence of reactions to a stimulus, and includes cognitive evaluations, subjective changes, autonomic and neural arousal, impulses to action, and behavior designed to have an effect upon the stimulus that initiated the complex sequence. ...Finally, there are eight basic reaction patterns that are systematically related to one another and that are prototype sources for all mixed emotions and other derivative states that may be observed in animals and humans.*

Scherer (1986) distinguishes different categories in a single emotion, like 'cold anger' and 'hot anger'. He summarizes that 'in reviewing the literature on the vocal expression of emotion, a discrepancy between reported high accuracy in vocal-auditory recognition and the lack of clear evidence for the acoustic differentiation of vocal expression is noted'. This still is a valid observation. In trying to come to a theoretical model of vocal affect expression, he found that the idea that emotion has to be seen as a process, and not as a steady state of the organism, becomes more common. Emotion is not a single response in one of the organism's subsystems (e.g., as physiological arousal or as subjective feeling or as motor expression), but rather emotion frames various components (e.g., physiological arousal *and* expression *and* feeling) in response to an *evaluation* of significant events in the organism's environment. According to Mozziconacci (1998) there are two main tendencies observable: one tendency is to consider emotions as discrete categories and another tendency is to view emotions as characterized by progressive, smooth transitions. In the first tendency a distinction is made between basic emotions and combinations of these basic ones. The latter tendency characterizes similarities and dissimilarities between emotions in terms of gradual distances on dimensions such as pleasant/unpleasant, novel/old, consistent/discrepant, and control/no control. Mozziconacci approaches her study by combining production and perception and first identifies parameters relevant for conveying emotion in speech. In chapter 2.3 we will review known acoustic cues.

Cowie et al. (2000) developed an instrument, based on the two dimensional

activation-evaluation space, a representation derived from psychology, to let users track the emotional content of a stimulus over time as they perceive it. The two dimensions respectively measure how dynamic the emotional state is and globally the positive or negative feeling associated with the state. Cowie et al. justify their approach, FEELTRACE, referring to research suggesting that the activation-evaluation is naturally circular, i.e. states which are at the limit of emotional intensity define a circle, with alert neutrality at the center. In their evaluation of the system they stress that their system fails to capture certain distinctions, like the distinction between fear and anger. This is to be expected, as they point out, since trying to capture emotion by projecting it onto only two dimensions will inevitably result in loss of information.

The thesis of Huisman (2004) forms a nice starting point when studying theories of emotion.

2.2 About speech production

In order to come up with a workable model for representing the speech signal, we need to have a decent understanding of the process of speech production. The speech wave is the response of the vocal tract filter system to one or more sound sources [16]. Applying this statement implies that the speech wave may be uniquely specified in terms of *source* and *filter* characteristics. It is widely described as a two-level process: the sound is initiated and it is filtered on the second level. This distinction between phases has its origin in the source-filter model of speech production [16, 25, 26, 54]. Actually, it would be more accurate to consider speech production as a three component system: a power supply (the lungs), an oscillator (the vocal folds), and a filter (vocal tract, i.e. supraglottal cavities), as remarked by Menzer (2004) and apparent in the work of, e.g., Schutte (1999).

In order to compose a framework to study speech production related issues, this section first briefly discusses the anatomy of the speech production apparatus, in order to become acquainted with some terminology[1]. Then aspects of speech production are discussed. The source-filter model itself is described in chapter 4 where the implementation of the vocoder is discussed.

2.2.1 Organs involved in speech production

We consider the physical system that gives rise to the speech signal, in order to visualize what we are talking about. Schutte (1999) gives a very good overview of the physiology of the production of voice. Figure 2.1 shows a schematized view of our speech (sound) production system. The anatomical structures involved in speech production can be divided into three groups, each with its specific role in the process of speech production. The glottis has a central position in all this. We then recognize the *subglottal system*, comprised of the lungs and their muscles, diaphragm,

[1] For convenience, translations to Dutch can be found in appendix B.

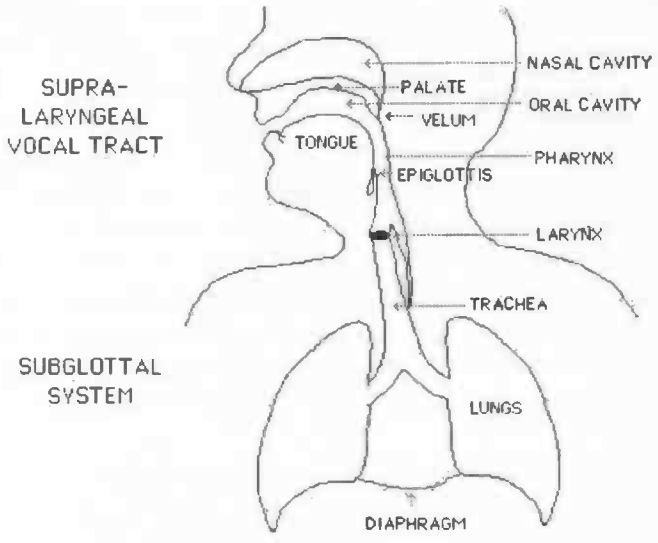


Figure 2.1: *Schema of the speech production system.*

and the trachea. Next there is the *glottal system*, which contains the larynx housing the vocal cords and glottis. Finally there is the *supraglottal system*[2], comprised of the structures above the vocal cords. These last structures can alter the shape of the upper vocal tract, notably the cavity of the mouth, enabling the realization of different sounds (e.g. timbre).

The space between the vocal folds is called the glottis (rima glottidis) [50]. The primary function of it is the closing of the trachea. A reflex will try to prevent food (when one is eating too voracious) or phlegm to enter the airway to the lungs. A forceful cough will blow unwanted materials out of the system. The secondary function of the glottis is producing voice. Driven by the lungs it is able to turn exhaled air into sound energy. The resulting pattern of vibration of the vocal folds is dependent on aerodynamic parameters of subglottal pressure from the lungs, and the divers adjustments and actions of the muscles in the larynx on top of the trachea, e.g. [50].

The larynx converts the steady flow of air produced by the subglottal system into a series of puffs, resulting in a *quasi-periodic*[3] sound wave. *Aperiodic* sounds are produced by allowing air to pass through the open glottis into the upper, supraglottal airway where localized turbulence can be produced at constrictions in the vocal tract. Normal respiration consists of an inhalation and an exhalation phase consuming about 3 seconds of time. When one speaks inspiration time is reduced substantial to about 0.5 seconds, whilst expiration can take up to 10 seconds [43].

[2]Also called the supralaryngeal vocal tract.

[3]A perfectly recurrent pattern in time is periodic. When there are small variations in period, amplitude, or both, the recurrent pattern is quasi-periodic [10]. Quasi-periodic waves are typical in nature.

The larynx is mainly composed of cartilage and above it the hyoid bone is situated, by which (via various muscles and ligaments) the larynx is connected to the jaw and skull. The larynx is composed of the thyroid, the cricoid, and the arytenoid cartilages. The vocal cords (or vocal folds –which is the same) are attached just beneath the laryngeal prominence (or more commonly known as the Adam's apple, the part of the thyroid creating the lump at the front of the neck), and the arytenoid cartilages. The arytenoid cartilages are three-sided pyramids and allow the vocal cords to be tensed, relaxed, or approximated (figure 2.2), thanks to the muscles in the larynx altering the inter vocal fold space and making the glottis more narrow or more wide. By this the human voice is able to produce its rich variety of sounds. This is of great importance since it is clear that the filtering characteristics of the supraglottal cavities alone can not account for this richness.

Fant (1970) states that the acoustic function of the vocal cords should not be regarded in analogy to vibrating membranes: they actually cause a modulation of the respiratory air stream, but do not generate sound oscillations of a significant magnitude by a direct conversion of mechanical vibrations to sound. A simple mechanical explanation to the vibrational mechanism can be given on the basis of the alternating force exerted on the vocal cords by the subglottal over-pressure in the closed state and by the negative pressure in the glottis in the open state due to the flow of air (ibid.). The air pressure in the trachea, which is virtually equal to the subglottal air pressure (denotes the pressure just below the vocal folds), is almost the same as in the lungs but the pressure above the glottis is nearly zero (i.e. like in the surrounding air). The latter sucking force, the Bernoulli effect, explains why the vocal folds can depart from an initial open state without muscle action (ibid). Through the lungs we can control the airflow in such a way that it is more constant and can be used longer (economics). For short utterances at normal loudness, normal expiration is sufficient. For louder or longer speech one needs to respire more deeply.

2.2.1. COROLLARY. *Subglottal pressure is one of the most important factors in speech production and primarily affects pitch and loudness.*

When a person is only breathing, the glottis is wide open, when voiced the glottis is almost closed. Figure 2.2 shows a schematized view of the glottis. The length of the vocal folds is dependent of gender and age: for females the folds are clearly shorter (13–17 mm) than for males (17–24 mm). For infants the folds are very short, like 5 mm. The length of the vocal folds, amongst other parameters, determines pitch. Pitch raises with vocal folds getting thinner and stiffer. It is determined that vowels have their own, intrinsic pitch (F_0): vowels like the /i/ and the /u/ have a higher pitch than the /a/. They differ about 4 to 25 Hertz and this phenomena is explained by skewing of the cricoid [40, 54]. F_0 (related to the perception of pitch) is inversely proportional to the vibrating mass and directly proportional to the tension of the folds. Assuming equal density (tissue density is constant for all phonation conditions [54]) and equal width of the folds, F_0 depends inversely on the length of

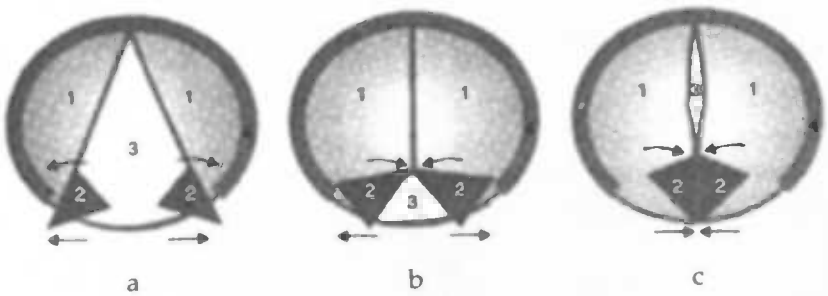


Figure 2.2: *(a) Glottis at voiceless sounds, (b) glottis while whispering, and (c) glottis during voiced sounds. Each subfigure shows the vocal folds (1), the arytenoid cartilages (2), and the glottis itself (3). From Rietveld and Van Heuven (2001)*

the vibrating part of the folds.

The variations of pitch of the human voice are possible due to the tension exhibited in the vocal folds. In Schutte (1999) an extensive overview is given of the structures of importance in the production of human sound. Worth mentioning is the fact that the microscopical anatomical construction of the glottis involves muscle fibers who, in most part, stretch spirally and are mutually interweaved, such that this results in tufts of fiber. This kind of muscular bundles are specific for human beings and make it possible to regulate the vocal folds very precisely.

The supraglottal system encompasses all parts playing a role in varying the cavities of the mouth. Those are the alveolen, de pharynx, the palatum and velum, the mandibula, the lips, the tongue, and the hyoid. By varying the cavities of the mouth, we are able to produce all kinds of sounds. In addition we may use the nasal cavity as an extra resonator to produce segments like a /m/ and a /n/.

2.2.2 Pattern of vibration of the vocal folds

Vibrating vocal folds move both in horizontal and vertical directions. On top of that, but virtually independent of it, the mucous membrane[4] exhibits an undulation or waving, that is able to move autonomously. This is depicted in Figure 2.3. The pattern of vibration changes as pitch changes. When pitch increases the amplitude of movement and undulation of the mucous membrane decrease. The time of closure during one complete cycle of vibration is related mostly to the intensity of the sound to be produced, and in lesser part to pitch. Figure 2.4 depicts the mechanism of closing time. As sound intensity increases the *open quotient*, OQ[5], decreases! At the same time the ratio of the speeds of the opening and closing movement in the open phase of the glottis cycle increases too. This means that the closing phase decreases. That is, the vocal folds close the glottis faster than they open it. As Menzer

[4]Mucous membranes are tissues that line body cavities or canals such as the throat, nose and mouth. Mucous membranes produce a thick, slippery liquid called mucus that protects the membranes and keeps them moist.

[5]Ratio between open phase and period.

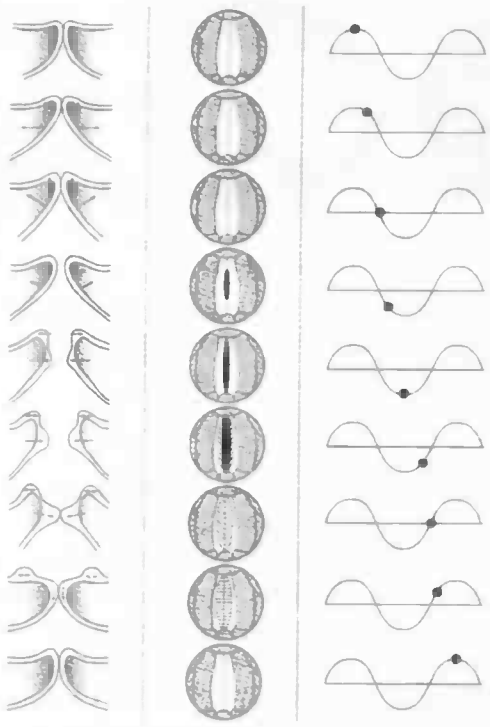


Figure 2.3: *Vibration pattern of the vocal folds during voicing. Left column shows a frontal cross-cut, the middle column shows the glottis as seen from above, and the third column shows the current phase of a complete glottis cycle. From Schutte (1999).*

(2004) describes it: 'therefore, in order to have energy provided to the vocal fold vibration by the glottal flow, it is necessary that the glottal flow is faster in the closing phase than in the opening phase.'

2.2.2. COROLLARY. *The changes in the ratios of the opening and the closing phase due to the closed phase, at increasing sound intensity, is an intrinsic property of the vibration pattern of the vocal folds.*

By means of the vibrating vocal folds and the, because of that, varied opening and closing of the glottis, a series of pulses is produced. These pulses exist of harmonic overtones with a regular decaying amplitude. Voice quality, at the level of the glottis, depends on the pattern of vibration. And this, of course, depends on many factors such as the thickness of the vocal folds, driving neural structures, et cetera.

Looking at normal functioning vocal folds, vibrations having a small amplitude and short closed phase (take a look at Figure 2.4 again) correspond with a more narrow spectrum; i.e. a spectrum containing fewer harmonics. At increasing subglottal pressure and voice intensity, a longer closed phase results. This, in turn, results in a wider spectrum, i.e. more overtones. The number of harmonics decreases when

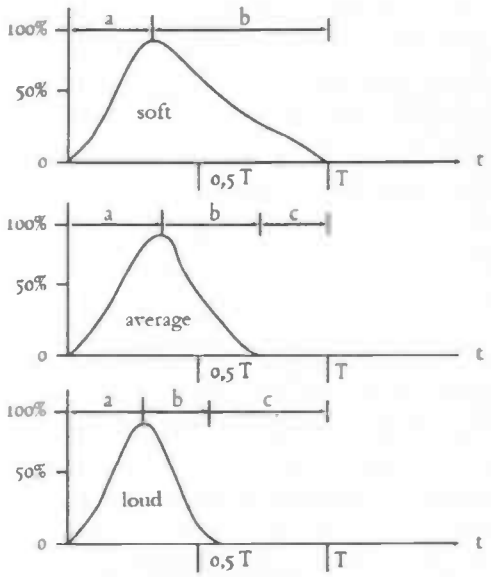


Figure 2.4: *Relationship between the open and closed phase of the glottis when sound intensity increases from soft to loud. (a = opening phase, b = closing phase, c = closed phase. T = one period.) From Schutte (1999).*

pitch increases, since the distance between the harmonics is equal to the frequency of the first harmonic.

Figure 2.5 depicts the complicated realization of voice. It shows five glottis cycles of the vowel /o/, at a frequency of 175 Hz. The vertical lines, A and A' mark the moments of glottis closure, at B the glottis opens. For healthy voices the supraglottal pressure is at a minimum at glottis closure, while at the same time the subglottal pressure is at its maximum. The glottis closes very abrupt and the resonance frequency of the subglottal cavity is visible in P_{sub} (in Fig. 2.5), having a frequency of about 550 Hz. In the supraglottal cavity a strong resonance at nearly the double frequency, here 350 Hz, is noticeable. Besides that various other resonances occur in the supraglottal cavities, they bring about the before mentioned formants; they determine our vowels.

2.2.3 Turning air into speech

There are three phases distinguishable in the process of producing speech sounds, as can be read in Rietveld and Van Heuven (2001): *initiation*, *phonation*, and *articulation*. These phases are briefly mentioned in the following paragraphs; the idea is to get a somewhat better understanding of the speech production process.

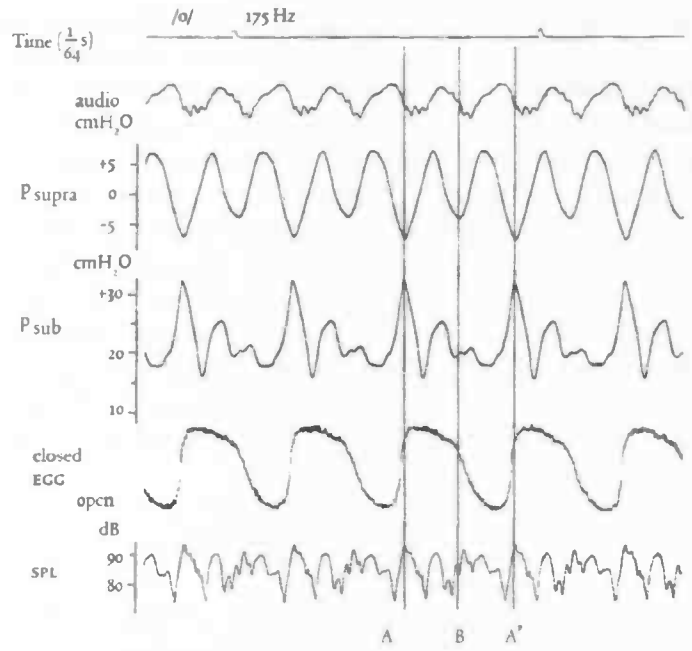


Figure 2.5: *Five glottal cycles of a vowel, /o/ spoken at 175 Hz. (Read text for explanation.) From Schutte (1999).*

Initiation

An airstream is initiated by the lungs and pushed via the trachea into the vocal tract (i.e. pulmonic egressive initiation) and it is the source of the sound production. Any constriction in the vocal tract (glottal or supraglottal) modifies the flow.

During inspiration, the lungs expand, causing the air to flow from the mouth to the lungs with the glottis relatively open. During expiration, the lungs contract, pushing the air from the lungs toward the mouth. Normally, and certainly for our purposes in this work, the production of sounds (phonation) occurs during expiration. The flow of air will be relatively small because of constrictions in the vocal tract and a nearly closed glottis. During normal breathing expiration the glottal area (take a look at Figure 2.2(c) again) is in he order of 1 cm^2, while during phonation the average glottal area is 0.05 to 0.1 cm^2.

Phonation

The larynx is used to transform an airstream into audible sounds. This process is called phonation and it is of special importance to perceived voice quality. The airflow through a narrowing glottis is transformed into short, periodic pulses. Depending on the volume of the airflow and the degree of constriction either *laminar* flow or *turbulent* airflow are effected. Turbulence occurs with higher airflow volumes and higher degrees of constriction. The vibrating of the focal folds is a repeating

process, which can occur at rates of, say, 80 to 500 cycles per second. The resulting, voiced speech[6] sound will show a certain distribution of higher amplitudes (air pressure). These occur at the times the vocal folds close under the Bernoulli effect. Shape, duration and amplitude of the pulses depend on muscular and aerodynamic factors. This process can be combined with other ways of generating sounds to create voiced fricatives or voiced stops. Phonation constitutes the fundamental set of voice quality parameters.

Articulation

The third phase in speech production is articulation. Articulation is the term used for all actions of the organs of the vocal tract that effect modifications of the signal generated by the voice source. This modification results in speech events which can be identified as vowels, consonants or other phonological units of a language. The transformation of the sounds generated during the phonation phase results from changing the supraglottal cavities into specific shape. We can divide these 'shaped' sounds into two classes: *vowels* and *consonants*.

The secondary function of articulation is to shape the paralinguistic[7] layer by 'coloring' and 'bleaching' the phonetic segments with the personality of the speaker.

The prosodic (and metrical) organization of an utterance also includes voice quality factors. The syllables in the chain of continuous speech are pronounced with different prominence. The prominence of a syllable involves the interaction of pitch, loudness, duration and articulatory quality. In most cases a more prominent syllable requires more muscular *effort* from the speaker. This muscular tension, but also changes in loudness, duration and articulation are perceived as a change in voice quality. Prosody is discussed in little more detail in section 2.3.1.

2.2.4 Vowels

We are concerned with the way vowels are produced, and what determines their *quality* –it is the focus of this work: synthesizing aggressive vowels. Schutte, referring to earlier work, claims that voice quality is not directly related to the quantities of lung capacity and lung volumina (like the amount of capacity used at respiration). However, the subglottal cavity and the supraglottal cavity are separated by the glottis, and the precise orchestration of the evolving pressures in those cavities, by the glottis, has a direct impact on the spectral content of the glottis pulse, and thus the quality of the voice.

[6]There are more differences between voiced and unvoiced sounds, other than the fact of vocal fold vibration.

[7]Paralinguistics is concerned with factors of how words are spoken, i.e. the volume, the intonation, the speed etc. Illustrative is that in intercultural communication paralinguistic differences can be responsible for, mostly subconscious or stereotyped, confusion. For example the notion that Americans are talking "too loud" is often interpreted in Europe as aggressive behavior or can be seen as a sign of uncultivated or tactless behavior. Likewise, the British way of speaking quietly might be understood as secretive by Americans. (Copied from Rietveld and Van Heuven (2001).)

Vowels are distinguished from other sounds by the fact that they are realized without the airflow being obstructed in the cavities of the mouth. That is, with normal speech, a *laminar* airflow is expected. The pulses generated by the vibrating vocal folds are 'refashioned' subsequently by the cavities of the mouth; the vocal tract selects, or tunes, a subset of frequencies produced by the glottis [16, 54]. These cavities can take all sorts of shapes by repositioning the various articulators. When producing vowels, the cavities of the mouth can be thought of –as a very simplified model– two coupled tubes. The tongue separates the tubes and a third tube, model for the nasal cavity, can get plugged in.

Figure 2.6 shows, for the vowels /i, e, ɪ, ɛ, a/, the position of three sensors on the tongue, giving us an idea of the shape of the tongue during realization of the vowels mentioned[8] The dotted line is there to compare the shape with the contour of the palatum. This gives us some idea of the principles of vowel production.

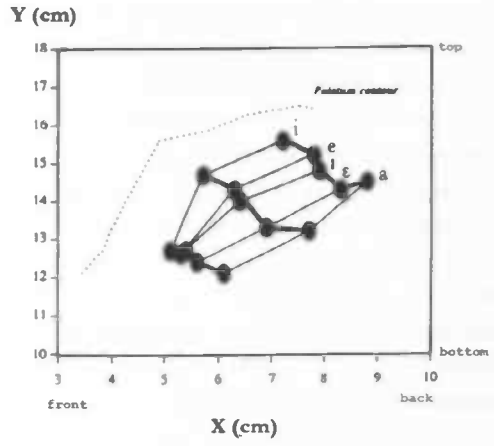


Figure 2.6: *Realization of the vowels /i, e, ɪ, ɛ, a/. The shape of the tongue is visualized. From Rietveld and Van Heuven (2001)*

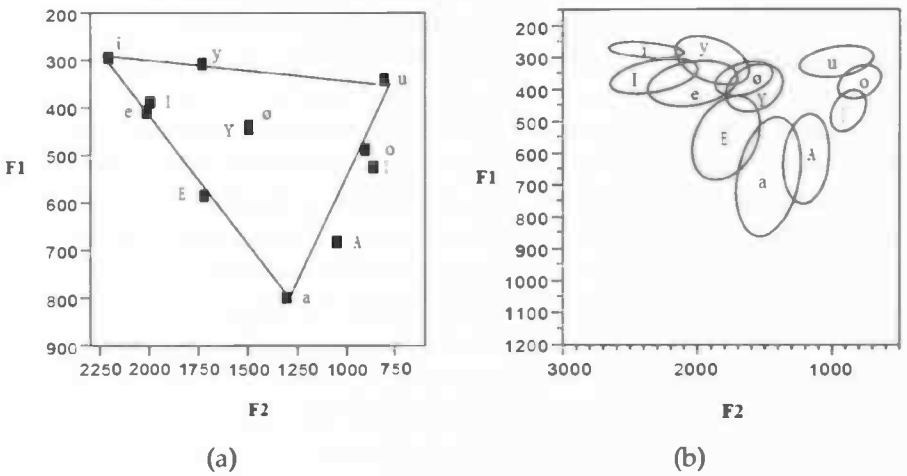
Definition A formant is a resonance of the vocal tract.

A formant is the result of resonances of the vocal tract. In the vocal tract energy is lost which results in broadening of the frequency spectrum of formants, i.e. formant bandwith. For simple acoustic sources this energy loss is proportional to the square of the frequency [54]. Every doubling of frequency therefore produces 6 dB more acoustic power. Energy loss must be understood in energy radiated from our lips and this, of course, is what we hear. When formant bandwidths increase, we perceptually tend to label a sound as *metallic*, for narrow bandwidths, and *muffled*,

[8] Here, the position of the tongue during the /e/ is a bit higher than for the /ɪ/. Would the position have been recordered somewhat later in time, then it would have been closer to the /i/. A slight slide – 'verglijding' (Dutch)– in articulation space is characteristic for three standard Dutch vowels. In the King's English these slides are even more profound. In other languages, e.g. French and German, the effect is absent for the vowels /e, ø, o/. (Copied from Rietveld and Van Heuven (2001).)

for broad bandwidths [54]. A research question would be what tendency formant bandwidths show, for vowels, in case of changing emotion.

It is common to classify vowels using the two, or three [39, 43], first formant frequencies, F_1 and F_2 (and F_3). This enables us to draw acoustic vowel charts, as depicted in Figure 2.7(a). Dispersion exists, due to speaker differences in genders, ages, e.g. Different languages or dialects bring, in some respects, different charts. Further it is assumed that we are dealing with normal, speaking voices; singing voices, for instance, would extend the range of formant frequencies. There have been many studies aiming at the description of vowels, like in [1, 42]. Since there is quite a variation in the absolute values of the formant frequencies between native speakers from different regions, it seems that not the absolute values of formant frequencies determine which vowel is perceived, but that it is the distance or perhaps their ratio[9] that is of importance, e.g. [39, 60]. The vowel chart in Figure 2.7(b) reflects this issue: it shows that there is a substantial spreading observable. When we



(a) (b)

Figure 2.7: *(a) Vowel chart for 12 Dutch vowels, based on the average formant frequencies of 50 Dutch men. Taken from Verhoeven and van Bael (2002); original work Pols et al. (1973), (b) Amongst men from the Dutch province Limburg, a substantial deviation from the mean formant frequencies is apparent. From Verhoeven and van Bael (2002). F_1 and F_2 are in Hertz.*

would draw triangles for, say, male and female habitants of a certain region, then we would have two triangles having approximately the same geometrical shape, but different sizes, the smaller one being the 'female' triangle, and the 'male' /i/-/a/ axes would have been moved closer to the /u/ corner [60].

It is not hard to imagine, acknowledging that the tongue plays an important role in the shape of the filters of the vocal tract, and that the mobility of the tongue is limited, mostly in the back of the mouth, that back vowels appear to be more stable than nonback vowels. The vowels /i/, /u/, and /a/ represent the extrema in tongue position and are called the *corner vowels*.

[9] In literature known as the formant-ratio theory.

We end this section with the well documented remark that, other things being equal, the average pitch of vowels shows a systematic correlation with vowel hight. That is, the higher the vowel the higher the pitch [40].

2.3 Acoustic cues

In this section we will take a look at what we are looking for: what are the cues? Ultimately, we want to know which cues determine speech to be perceived as aggression. Question to ask is, among many others, how much overlap a certain cue shows for (related) emotions?

2.3.1 Aspects of prosody

Prosody, or the way things are spoken, is an extremely important part of the speech message. Changing the placement of emphasis in a sentence can change the meaning of a word, and this emphasis might be revealed as a change in pitch, volume, voice quality, or timing.

In modeling speech, one can distinguish global and local properties [35]. Among the global properties are the overall pitch range typical of a given speaker, the actual pitch range used in the utterance, the amount of declination, the rate of speech, rhythm variations, and so on. Although such properties are essential for simulating emotions or speaking styles, one makes abstraction of them when interpreting the linguistic functions of intonation (such as prosodic boundaries, prosodic organization and focus). The underlying assumption is that a (structural) pitch pattern (configuration, contour) may be modulated by global parameters in order to express information carried by the pitch pattern and by global properties simultaneously (ibid).

Prosody is not used in our current experiment, although we recognize it's importance. The one thing we did is to apply a pitch contour when we synthesized the data for the experiment. Not doing so would, possibly, make it too obvious the data were synthesized vowels.

2.3.2 Which cues?

Although humans are very well equipped to classify a rather broad range of emotions, a definition of any of those emotions is an other matter. Hagmüler et al. (2004) have concluded that the human voice as a tool for stress observation shows a high potential. They state that there is a lot of research carried out by either psychologists or linguists, who have verified statistical significance of vowel cues for voice stress observation. In Murray et al. (1996) some definitions of stress are given which reside in literature. In their article they come to the conclusion that the effect of stress on speech is poorly understood due to its complexity: it is not clear how changes in a perceived speech signal relate back to the stressors. They end with saying that there

are many proposed definitions of stress, models of stress, stressors, strain effects, and how to measure all of these, but none have unanimous support.

Browsing literature one learns that the most frequently used cues for observation of emotions in human speech are the fundamental frequency or pitch. It is considered to be dependently related to human arousal[10]. In Alonso et al. (2005) classical characteristics have been divided into five groups depending on the physical phenomenon that each parameter quantifies, namely quantifying the variation in amplitude (shimmer), the presence of unvoiced frames, the absence of wealth spectral (Hitter), the presence of noise, and the regularity and periodicity of the waveform of a sustained voiced speech sound. To measure acoustic cues of emotions (i.e. feature extraction) with high ergotropic arousal, e.g. aggression, often loudness, voice quality, and pitch are used.

Scherer (1986) describes specific predictions to the changes in acoustic parameters resulting from changing physiological responses characterizing different emotional states. His predictions show similarities with the acoustical effects of the *Lombard reflex*. Lombard was the first to examine the influence of raising of the voice on acoustical properties of speech. The Lombard reflex has an effect on the loudness of speech and on the quality of voice: 'In the presence of noise, speech is masked, and its production is modified by what is called the Lombard effect. The Lombard effect is the reflex that takes place when a speaker modifies his vocal effort while speaking in the presence of noise', according to Junqua (1993). *Voice quality* is related to the amount of distinguishability of the harmonics in a signal, that is, higher voice quality entails better distinction of the separate harmonics [20].

Acoustical cues related to *pitch* are F_0, bandwidth of F_0, contour and the amount of fluctuations of F_0-contour, called shimmer and jitter in Huisman (2004). Acoustical cues related to *loudness* are average energy, relative amount of energy, and fluctuations in energy. A speech signal can be divided into several frequency bands and then cues are estimated for these ranges, e.g. Banse and Scherer (1996). Acoustical cues related to high *voice quality* are visualized as clear peaks of harmonic frequencies in a spectrogram. Cues are extrema (estimated maxima and minima) in the energy spectrum.

Huisman states that in his research mainly spectral cues are examined. But it is expected that there is useful information in the temporal dimension of the data. Leaky integration in the SI model of the cochlea, as explained in Andringa (2002), which was used for measurements and analysis, was likely to diminish the effects of jitter and shimmer in Huisman's research.

Toivanen et al. (2003) present 41 prosodic parameters measured form the speech signal. These partially overlap the before mentioned ones. Junqua (1993) discusses the *Lombard reflex*.

[10] Arousal is a physiological and psychological state involving the activation of the reticular activating system in the brain stem, the autonomic nervous system and the endocrine system, leading to increased heart rate and blood pressure and a condition of alertness and readiness to respond. It is a crucial process in motivating certain behaviors, such as the fight or flight response and sexual activity. It is also thought to be crucial in emotion, and has been an important aspect of theories of emotion.

Next acoustic parameters as found in literature, e.g., Banse and Scherer (1996) and Klasmeyer (2000), are given. This should give an idea of the perceptual cues in use today.

- Fundamental frequency, F_0: mean, standard deviation, 25^{th} percentile, 75^{th} percentile, range of F_0 and ΔF_0, minimal value, maximal value;
- Energy: mean, standard deviation, energy of high frequencies, word energy, energy of syllables;
- Speech rate: duration of fricatives, plosives, sonorants, vowels, duration of syllables, phonemes, words, pauses;
- Voiced long-term average spectrum: 125-200 Hz, 200-300 Hz, 300-500 Hz, 500-600 Hz, 600-800 Hz, 800-1000 Hz, 1000-1600 Hz, 1600-5000 Hz, 5000-8000 Hz;
- Unvoiced long-term average spectrum: 125-250 Hz, 250-400 Hz, 400-500 Hz, 500-1000 Hz, 1000-1600 Hz, 1600-2500 Hz, 2500-4000 Hz, 4000-5000Hz, 5000-8000 Hz;
- Hammarberg index, which measures the difference of energy maxima in the 0-2 kHz band and the 2-5kHz band in the voiced part of the utterance;
- Slope of spectral energy above 1000 Hz, proportion of voiced energy up to 500 Hz, proportion of voiced energy up to 1000 Hz;
- Hilbert Envelope in different Frequency Bands, i.e. the distribution of noise in the voiced signal.

Having these acoustic cues we next want to know how to map emotions and cues onto each other. There is a vast amount of literature available on this subject. We like to mention the work of Schröder et al. (2001) and Schröder (2004). Their work is also related to the FEELTRACE tool [11], mentioned in section 2.1. This tool projects emotions on a activation and a evaluation dimension. As an example Schröder et al. explain that the emotions anger and fear are very close on the activation and evaluation dimension. Fear and anger are similar in pitch average, pitch range, speech rate and articulation (ibid). Schröder et al. continue saying that fear differs from anger in that pitch changes are not steeper than for neutral, the speech rate is even faster, the intensity is only normal, and voicing is irregular.

We have to refer to literature, if one is interested in more detail, and only mention that there indeed is considerable overlap between emotions for current cues. This fact calls for –as we see it– a change of perspective; in this work focus is on speech production and we tried to define a small set of cues/parameters related to it. These cues will be defined in chapter 6 when we outline the method of our experiment. It is our belief that effects measured on the side of a receiver, perceptual or using cochleograms [3], may relate back to more than one source at the production side.

Thus, to be able to come to a one-to-one mapping of cause and effect, we have to look closer at the speech production. We would like to be able to predict perceptual changes based on what happens with our vocal folds, for instance. When glottal pulses become shorter in time, due to some emotional change of a speaker, we can predict that more energy of higher frequencies will be measured. So, instead of post hoc analysis of a speech signal, we, hopefully, are directed towards cues by physiological evidences. In this work we test if such an approach is viable by synthesizing our own data and letting subjects label it. We further know which supposed cues we have put into the signal and are therefore able to assign cause and effect mappings. In genuine data we had to rely on pattern recognition, as it were, now we can test hypotheses. Of course, a vocoder does not eliminate the need to analyze recordered speech. Our assumption is that both approaches need each other, that is, they, at least, may profit from each other and can serve as a boost for one another.

2.4 Nonlinear analysis

This section is introduced to reflect interesting progress in recent work. It is founded on the idea that linear modeling cannot account for all phenomena found in speech processes, e.g. Little et al. (2006), Zhou et al. (2001). These phenomena, however, might be of particular interest when looking for cues determining aggressive speech, and moreover, theoretical and experimental evidence is becoming strong [32]. Asogawa and Akamatsu (1999) suggested in their paper that the vowel is the product of a nonlinear system and, as is always the case with nonlinear systems, they explain, the dynamics of the system exhibit such peculiarities as bifurcation, lock-in of frequency, intermittency, and chaotic behavior, non of which occur in linear systems.

2.4.1 Something about bifurcations

In Menzer (2004) transient behavior in the vibration of the human vocal folds is studied. In particular *pitch breaks* (sudden changes in the fundamental frequency) with a non-integer frequency ratio were found to be interesting because this case is different from the classic *period-doubling* scenario. Menzer writes:

> *A new physical model for the vocal folds was [also] developed, with the aim of keeping the number of state variables as low as possible. The result is a third order system having the contact area and the glottal airflow as state variables. In terms of the number of state variables this system is in the same class as a one-mass model driven by the glottal flow. However, it has more features than are usually found in a one-mass model. It also simulates the zipper-like opening and closing of the folds and takes into account a deformation of the vocal fold tissue.*

Menzer devoted particular attention to the development of a new model of the vocal folds that takes into account their zipper-like movement, as he calls it, see

section 2.2.2, and the fact that vocal fold tissue is not rigid. This model has only three state variables, including contact area and glottal flow.

Further is noted that 'an application of bifurcating nonlinear models could be to use them to drive real-time voice synthesis. This may contribute to a more natural sound. However, it must be considered that *much of the naturalness of a sound has little to do with the vibration model itself, but with the way it is controlled.*'. The latter fact is acknowledged by us, see also section 6.1.

A real important remark is made by Menzer. It seems as if the voice source coming from the physical model produces less "buzziness" than found for instance in the Liljencrants-Fant (LF) model. Buzziness is an unwanted artifact of voice synthesizers. The LF model has this problem, probably because its derivative is discontinuous.

It is hypothesized that pitch breaks are due to a constriction of the airflow above the vocal folds. Depending on the speaker, period doubling was found very often. This contrasting the findings of Schutte, who we spoke about the subject, and who claimed that period doubling is to be expected of not much use due to the amount of variance found between people. Indeed, Menzer clearly stated that it depended on he speaker. Period doubling is characterized by one peak out of two decreasing or increasing in amplitude. This behavior is commonly found in well-studied nonlinear systems such as the Colpitts oscillator. Subharmonic pitch breaks are interesting in this context, according to Menzer, for several reasons. On the one hand, depending on the speaker they can occur relatively often in natural speech. On the other hand, period doubling is one of the most studied and well-known phenomena related to nonlinear systems. The sound of the period doubling is mainly perceived as a change in fundamental frequency. Menzer found that instead of creating lower subharmonics in a series of $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$, times the fundamental frequency, the vocal folds are able to create subharmonic ratios that are not a power of 2. He claims that ratios of $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ and $\frac{1}{5}$ times the fundamental frequency have been observed. Certain prerequisites have to be met before these period changes can emerge.

Classic period doubling

Period doubling is a well studied phenomena in in nonlinear systems [34, 53], or chaos theory in more popular terms. What happens is that when changing a parameter of certain systems (e.g. the logistic map) at certain parameter values the period of the observed signal doubles, giving rise to a sequence of doubling periods: $\{T, 2T, 4T, 8T, \ldots, 2^i T, \ldots\}$. At some point we speak of chaotic behavior. But within chaotic regions usually periodic "windows" are found. Figure 2.8 shows the route to chaos for the logistic mapping, $x_{n+1} \rightarrow kx_n(1 - x_n)$, when parameter k is increased from 2 to 4. The bifurcation diagram nicely shows the forking of the possible periods of stable orbits from 1 to 2 to 4 to 8, and so on. The vertical bands are the periodic windows.

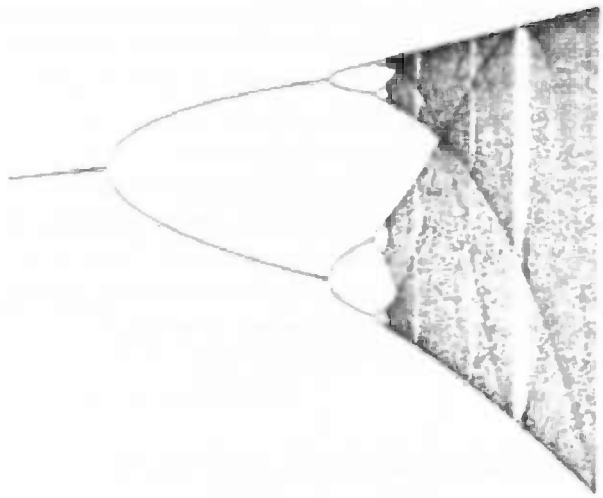


Figure 2.8: *Bifurcation diagram for the logistic mapping ($x_{n+1} \rightarrow kx_n(1 - x_n)$). On the horizontal axis the constant k increases from 2 to 4, vertically the state x is shown. Period doubling is followed by the growth of chaotic bands.*

2.4.2 Dynamic modeling

Nonlinearities in speech is treated in many articles, e.g. [30, 31, 32]. Maragos et al. (2002) summarize their on-going work on structures caused by modulation and turbulence phenomena, using the theories of modulation, fractals and chaos. A novel approach for vowel classification by analyzing the dynamics of speech production in a reconstructed phase space is presented in Liu et al. (2003). Their drive is the fact that conventional linear spectral methods cannot properly model nonlinear correlation within the signal, and therefore, they argue, methods that preserve nonlinearities may be able to achieve high classification accuracy; preliminary results clearly indicate the potential of dynamics analysis for speech processing, also in the context of stress detection and classification. By using the fact that one revolution of a three dimensional reconstruction of the speech signal is equal to one pitch period (attractor reconstruction in state space, Poincaré maps, e.g. [53]), Mann and McLaughlin (1998) derived a new algorithm for epoch marking and describe their technique as promising however not to be taken as a competitor to existing techniques. They merely wanted to demonstrate practical possibilities that nonlinear signal processing has to offer.

Besides using nonlinear techniques to analyze the speech signal, researchers are investigating methods to model the voice producing element itself [12, 13, 34]. De Vries et al. (2002) and De Vries et al. (2003) use numerical models of the vocal folds based on the two-mass models of the vocal folds. They couple it to a model of the

glottal airflow based on the incompressible Navier-Stokes equations (for which computation is still heavy). Results are compared against Bernouilli-based models; De Vries et al. explain that the use of the Bernouilli equation is allowed when, among other restrictions, the glottal airflow is assumed to be steady and laminar. This will not be the case when aggression is displayed.

The next section introduces the Teager Energy Operator. It is discussed separately because it will be discussed in more detail. We would like to test it's use for aggression detection and classification in future work. There was a lack of time to do this in this work, unfortunately. Currently, researchers are optimizing and extending the energy operator, using it in the context of emotion recognition, e.g. Zhou et al. (2001).

2.5 Speech modulations

Very often linear models are used to analyze speech signals. Although this approach seems to work very well for a broad range of applications, there certainly are nonlinear effects associated with speech production. These effects might contain information not apparent in normal analysis methods. Evidences for speech modulations are found in several experimental and theoretical works [33]. Most of these evidences are centered around ideas of analyzing the dynamics of speech production using concepts from fluid dynamics to study properties of the speech airflow (ibid).

Shadle et al. (1999) state that the evidence points toward the existence of a vortex train during and caused by phonation, and significant sound generation due to the interaction of that train with tract boundaries; these findings indicate that the models on which inverse filtering[11] are based have been overgeneralized. More recently Little et al. (2006) showed that the linear prediction analysis cannot account for all the dynamical structure in speech. This does not mean that the classical assumptions can be ruled out. It could mean, however, that it might be useful to study some nonlinear properties of speech as to find out whether they provide us with better predictors of the emotional content of speech.

Maragos et al. (1993) described a nonlinear differential operator that can detect modulation patterns in speech signals. A great advantage is the fact that such *energy separating algorithms* (ESA) can have a very low computational complexity, are efficient and have an instantaneously-adapting nature. There have been many implementations of ESAs, here we discuss a discrete ESA based on the Teager-Kaiser energy operator [32, 33]. Maragos et al. summarize the promises of discrete ESA: (i) it yields very small errors for amplitude and frequency demodulation, (ii) it has an extremely low computational complexity, (iii) it has an excellent time resolution,

[11] Inverse filtering is a method to reconstruct the shape of the glottal pulse train. It assumes the filtering characteristic of the supra glottal vocal tract is known, which makes it possible to apply inverse filtering using the convolution theorem. There are, however, several difficulties involved in this approach.

almost instantaneous, (iv) it is less computationally complex and has better time resolution than other classical demodulation approaches, (v) it can track the true physical energy of the acoustic source, and (vi) it can detect transient events which can be useful for, e.g., detecting plosive sounds.

Dimitriadis and Maragos (2001) compared some algorithms and developed a spline based approach which yields better results under noisy circumstances. In the next section an AM-FM signal is described and the Teager-Kaiser energy operator is introduced. It is based on the work of Dimitriadis and Maragos.

2.5.1 Demodulation

We are discussing signals which are modulated in amplitude (AM) and in frequency (FM). Such signals can be represented as

$$x(t) = a(t)\cos(\phi(t)) = a(t)cos(\int_0^t \omega(\tau)d\tau + \phi(0)), \tag{2.1}$$

where $\omega(t) = d\phi/dt$. The total speech signal is supposed to be the superposition of such signals, one for each formant. Here $a(t)$ is the instantaneous amplitude signal and $\omega(t)$ is the instantaneous angular frequency representing the time-varying formant signal. The short-time formant frequency average $\omega_c = (1/T)\int_0^T \omega(t)dt$, where T is in the order of a pitch period, is viewed as the carrier frequency of the AM-FM signal. The classical linear model of speech views a formant frequency as constant, i.e., equal to ω_c, over a short time period (e.g., 10-30 ms). However, the AM-FM model can both yield the average ω_c and provide additional information about the formant's instantaneous frequency deviation $\omega(t) - \omega_c$ and its amplitude intensity $|a(t)|$. To isolate a single resonance from the original speech signal, band-pass filtering is first applied around estimates of formant center frequencies.

We are interested in $a(t)$ and $\omega(t)$. If we rewrite $x(t)$ we are able to calculate amplitude, frequency and phase:

$$\begin{aligned} x(t) &= a(t)e^{\imath\phi(t)} \\ &= x_r(t) + \imath x_i(t), \end{aligned}$$

x_r and x_i being the real and imaginary parts of the signal $x(t)$. Now it is possible to write down the equations for the unknowns:

$$\begin{aligned} a(t) &= |x(t)| &&= \sqrt{x_r(t)^2 + x_i(t)^2} \\ \phi(t) &= \arg(x(t)) &&= \arctan(x_i(t)/x_r(t)) \\ \omega(t) &= \operatorname{im}(\dot{x}(t)/x(t)) &&= \frac{\dot{x}_i(t)x_r(t) - x_i(t)\dot{x}_r(t)}{x_i(t)^2 + x_r(t)^2}. \end{aligned}$$

The definition of the Teager Energy operator is

Definition

$$\Psi(x(t)) = \dot{x}(t)^2 - x(t)\ddot{x}(t), \tag{2.2}$$

where $\dot{x}(t) = dx(t)/dt$. This defines the operator for *continuous* signals. If we apply the (continuous) Teager operator to a signal $x(t) = A\cos(\omega t)$, a signal without modulations, equation 2.2 yields:

$$\begin{aligned}\Psi(x(t)) &= (-A\omega\sin(\omega t))^2 - A\cos(\omega t)\left(-\omega^2 A\cos(\omega t)\right) \\ &= A^2\omega^2\left(\sin^2(\omega t) + \cos^2(\omega t)\right) \\ &= A^2\omega^2, \end{aligned} \tag{2.3}$$

which is a constant. For the discrete case things are a bit more complicated. Let's define a discrete signal as

$$x[n] = A\cos(\Omega n + \phi), \tag{2.4}$$

where $\Omega = 2\pi f/F_s$, F_s is the sampling frequency and f the sampled analog frequency. Since this equation has three unknowns, we can determine them by solving a set of three of these formulas, each having a different n. Thus, we could set up three equations as follows

$$\begin{aligned} x[n] &= A\cos(\Omega n + \phi) \\ x[n-1] &= A\cos(\Omega(n-1) + \phi) \\ x[n+1] &= A\cos(\Omega(n+1) + \phi). \end{aligned}$$

Using trigonometry[12] we get

$$x[n-1]x[n+1] = A^2\cos^2(\Omega + \phi) - A^2\sin^2(\Omega)$$

and we recognize that the first term on the right-hand side of the equation equals x_n^2, so we can write

$$A^2\sin^2(\Omega) = x[n]^2 - x[n-1]x[n+1]. \tag{2.5}$$

From here we can derive the Teager energy operator (TEO) assuming the following: equation 2.5 has a unique solution if $0 \leq \Omega < \pi/2$. In Kvedalen (2003) we learn that this will be the case if Ω is less than one quarter of the sampling frequency. If we use the fact that for small Ω the approximation $\sin(\Omega) \approx \Omega$ holds, then we end up with the discrete TEO, since $A^2\Omega^2 \propto$ energy:

$$\Psi[x[n]] = x[n]^2 - x[n-1]x[n+1]. \tag{2.6}$$

Thus the TEO embodies an approximation, but the approximation error is less than 11% if $\Omega < \pi/4$, which is easily verified. Summarized, we can calculate, with small error, the instantaneous energy of a signal $x[n]$ by using $\Psi[x[n]]$.

But this is not it, applying the operator on a modulated signal perhaps is a bit cumbersome. Let's examine the result of the Teager operator applied to a both in

[12] The following identities can be used:
$\cos(\alpha+\beta)\cos(\alpha-\beta) = \frac{1}{2}(\cos(2\alpha) + \cos(2\beta))$ and $\cos(2\alpha) = 2\cos^2(\alpha) - 1 = 1 - 2\sin^2(\alpha)$.

amplitude and frequency modulated signal. We will not derive the result but instead copy it from literature, in casu Dimitriadis and Maragos:

$$\Psi\left(a(t)\cos(\phi(t))\right) = \left(a(t)\phi(t)\right)^2 + \underbrace{\frac{1}{2}a(t)^2\ddot{\phi}(t)\sin(2\phi(t))}_{FM} + \underbrace{\cos^2(\phi(t))\Psi(a(t))}_{AM}. \quad (2.7)$$

In equation 2.3 we had the result for an unmodulated signal (we are still talking about sinusoid signals). If we apply the operator on the differentiated signal, $\dot{x}(t)$ we obtain:

$$\begin{aligned}\Psi\left(\dot{x}(t)\right) &= \Psi\left(-A\sin(\omega t)\right) \\ &= A^2\omega^4\cos^2(\omega t) - \left(-A\omega\sin(\omega t)\right)\left(A\omega^3\sin(\omega t)\right) \\ &= A^2\omega^4.\end{aligned} \quad (2.8)$$

Combining equations 2.3 and 2.8 gives us the following nice results:

$$\omega(t) \approx \sqrt{\frac{\Psi\left(\dot{x}(t)\right)}{\Psi\left(x(t)\right)}} \quad (2.9)$$

$$|A(t)| \approx \frac{\Psi\left(x(t)\right)}{\sqrt{\Psi\left(\dot{x}(t)\right)}}. \quad (2.10)$$

These are *approximations* if the Teager operator is used on modulations. But, the errors are small when the FM and AM terms in equation 2.7 are small compared to the term $a^2(t)\phi^2(t)$.

These last results can be used for continuous signals. In our case we like to have comparable equations to be used on discrete signals. Now, there exist more versions of the Teager operator, because there are many ways to approximate the continuous derivatives. We are not going to give the possibilities here, but instead some results from literature are given [14] which are discussed extensively in Maragos et al. (1993). Approximating $\dot{x}$ by the two-sampled differences the following approximations are derived (note that in order to unclutter the formulae $x[\bullet] \equiv x_\bullet$):

$$\Omega[n] \approx \arccos\left(1 - \frac{\Psi\left[x_n - x_{n-1}\right]}{2\Psi\left[x_n\right]}\right) \quad (2.11)$$

$$|A[n]| \approx \sqrt{\frac{\Psi\left[x_n\right]}{1 - \left(1 - \frac{\Psi\left[x_n - x_{n-1}\right]}{2\Psi\left[x_n\right]}\right)^2}}. \quad (2.12)$$

The algorithm given in the last two equations is called DESA-1a[13]; algorithms DESA-1a, DESA-1 and DESA-2 also exist.

[13] DESA stands for Discrete ESA, the '1' indicates the two sample difference, and the 'a' means that an asymetric derivative has been used.

2.5.2 Noise

Besides the advantages of the Teager energy operator, there are some disadvantages. The main problem is that discrete differentiators are very sensitive to noise. In addition, higher frequencies appear in the operator. Recent literature is coming up with extended versions of the operator in the context of emotion detection, e.g. Zhou et al. (2001), which have improved properties. We discuss the operator applied on a noisy signal $s(t)$, where the noise is assumed to be white (Gaussian) noise, $n(t)$, with zero mean and variance σ^2. We can define our noisy signal as

$$s(t) = x(t) + n(t),$$

where $x(t)$ is the noise-free signal. What we are looking for is the expected value of the noisy signal. The expected value of a signal [19] is defined by

$$E[x] = \int_{-\infty}^{\infty} x(\alpha) f_x(\alpha) d\alpha,$$

where f is the probability density function (pdf) of x. Applying the expectation operator on the Teager operator yields [19, 27]

$$\begin{aligned} E\{\Psi(s)\} &= E\{\dot{s}^2 - s\ddot{s}\} \\ &= E\{\dot{s}^2\} - E\{s\ddot{s}\} \\ &= E\{[\dot{x} + \dot{n}]^2\} - E\{[x + n][\ddot{x} + \ddot{n}]\} \\ &= E\{\dot{x}^2 + 2\dot{x}\dot{n} + \dot{n}^2\} - E\{x\ddot{x} + x\ddot{n} + n\ddot{x} + n\ddot{n}\}. \end{aligned}$$

Realizing that n is independent of x we can remove all cross terms in the last equation, giving us

$$\begin{aligned} E\{\Psi(s)\} &= E\{\dot{x}^2 + \dot{n}^2\} - E\{x\ddot{x} + n\ddot{n}\} \\ &= E\{\dot{x}^2 - x\ddot{x}\} + E\{\dot{n}^2 - n\ddot{n}\} \\ &= E\{\Psi[x]\} + E\{\Psi[n]\}. \end{aligned} \tag{2.13}$$

In the discrete case we may write [27]

$$E\{\Psi[s]\} = E\{\Psi[x]\} + \sigma^2. \tag{2.14}$$

This result is important when implementing the Teager operator in a practical application. It tells us that the Teager energy estimates will be biased by the variance of the noise signal. (The amplitude and frequency estimates will be skewed.)

A future practical approach

For our purposes we want to apply the operator on formants. When we have an estimate for the average formant frequency we can device a bandpass filter and try

to minimize the influence of noise. In literature the *Gabor filter* is often used, e.g. [36]. The Gabor filter is described as

$$h(t) = e^{-a^2t^2}\cos(\omega_c t) \tag{2.15}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2a}\left(e^{-\frac{(\omega-\omega_c)^2}{4a^2}} + e^{-\frac{(\omega+\omega_c)^2}{4a^2}}\right), \tag{2.16}$$

where ω_c is the center angle frequency and a is the bandwith of the filter. Gabor filters are defined by harmonic functions modulated by a Gaussian distribution. They are advantageous because of their sideband properties. In this work there has been no time to dig into this matter, unfortunately, but the next lines give –a very premature– approach to use the TEO in jitter and shimmer (respectively frequency and amplitude modulations) detection. It should employ the fact that fundamental frequency, F_0, is expected in a frequency band, B_{F_0}. Evidence is available in the harmonics, $n \times F_0$. By means of *evidence feedback* the search area is narrowed until F_0 is locked in. Multiple Gabor filters can be used, each having a center frequency equal to a harmonic. A supervisor module will track the (slowly) changing pitch and raise a flag allowing for demodulation.

This method is expected to work on voiced speech segments and it will probably be advantageous to combine it with other methods (onset detection, et cetera).

Chapter 3

Research objectives

An idea that is developed and put into action is more important than an idea that exists only as an idea.

Buddha 563–483 B.C.

People are very robust emotion detectors and classifiers. They function under very different circumstances and moreover, they are able to discern the very subtle gradations of an emotion. One and the same word pronounced in an angry, or fearful manner sounds different. A lot of research is still going on to find those aspects of speech which determine whether and which emotion is present is the verbal message. Obviously, information in spoken language is not coded in the verbal component alone, but also in a nonverbal components (e.g. facial expression). Using this multi-modality makes it clear for an 'audience' what the real emotional state of the speaker is. This is of course true, but humans are very skilled in, almost instantaneously, valuing speech on a broad spectrum of emotions, one of which is aggression.

The objective of this thesis can be formulated as:

The aim of this research is to develop, implement, and evaluate an applicable aggression synthesis method (vocoder).

We have to, as far as the implementation is concerned, determine which method best adheres to the methods and instruments used at `Sound Intelligence`. It is in their interest to develop a vocoder which can be used to manipulate spectral aspects of vowels, to begin with. Perceptual testing of these synthesized vowels, using test subjects, will hopefully bring us a bit closer to the understanding of the human capability of aggression detection and classification.

In literature numerous acoustical cues are formulated that play a role in emotion perception. The research question is closely related to the objective of this thesis and is formulated as follows:

Which spectral features caused by the speech production system that can be resolved from the signal can best be used to classify aggression?

This formulation is based on the fact that we, i.e. Sound Intelligence, use a model of the human cochlea to process a sound signal. However, the cochlea is only one part of the human speech (sound) processing system. A substantial amount of processing is taking place in the cortex and other area's of the brain. Also, humans have

a lifetime of experience enabling them to use all kinds of knowledge. So actually, a very important question is how far a cochlea model will bring us? This question will not be answered in this work, obviously, but directions for future work will be suggested.

The research question and research objective enable us to define a few guidelines that can help us to reach our goal:

Action *Determine which synthesis method is applicable.*

Action *Define a small set of acoustical cues people might use in aggression detection and classification, focussing on the speech production.*

Action *Use the vocoder to test the cues in a hearing test.*

Action *Determine, by means of a database of synthetic vowels, generated by the vocoder, whether the vocoder is scientific justified in resolving the research question.*

3.1 Scientific relevance

Sound and speech recognition is a very important research area in artificial intelligence. Knowledge of the relation between emotions and acoustical features in speech is of importance in speech recognition. One, of the many possible, applications can be in Human-Machine interaction. Currently, Sound Intelligence is working on the development of the next generation of aggression detectors. Those systems are aimed at not only detecting aggression, but also classifying it. This work is closely related to their intentions.

My wish would be that the results of this work contribute, however small, in the understanding of the structures underlying aggression communication. And, hopefully, the vocoder will prove to be an useful tool in this ongoing quest.

Chapter 4

Source-filter modeling the vocoder

Only speaking machines are capable of producing a perfectly monotonic pitch.

Gunnar Fant

Speech synthesis techniques may be divided into three categories: concatenative, articulative, and formant speech synthesis. In concatenative speech synthesis a database is used to store segments of existing speech. This kind of speech synthesis uses text as an input, divides the text into minimal parts which are represented by the segments stored in the database. These segments are looked up and concatenated, using rules, to form the verbal equivalent of the text. The resulting speech may suffer of audible glitches resulting from the gluing together of the segments, implying discontinuities between them. Articulative speech synthesis emulates the speech organs, while formant speech synthesis simulates formant frequencies.

The emulation is done by mathematical computational models and produces, in principle, the most natural speech sound of the three techniques. But, obviously, it is also the most difficult approach [28]. The method involves models of the vocal cords and the articulators, often modeled as small tube sections. The vocal cords may be modeled using the two-mass model, which may be coupled to a model of the voice-producing element described by the incompressible Navier-Stokes equations, using techniques as the finite element method, et cetera, e.g. [8, 12, 13]. Formant speech synthesis reconstructs the formant characteristics by defining a set of resonators, a voicing source and a noise source. Effects of the nasal tract can be simulated by plugging in antiresonators. The same goes for plosives and fricatives. In formant synthesis the set of parameters (or rules) controlling the frequency and amplitude of the formants can be large [28]. Of course, there is also much to be said about the properties of the sources. The problems become apparent listening to the results: often there is some lack of naturalness.

This chapter goes into more detail concerning the filter part of the model, the voicing source is discussed in chapter 5 and follows the Liljencrants-Fant model. First the source-filter model is introduced, then the implementation of the vocoder is explained. Appendix C proposes a possible user interface (GUI).

4.1 The source-filter model

The basic assumption of the model is that the source signal produced at the glottal level is linearly filtered through the vocal tract. The resulting sound is emitted to the

surrounding air through radiation at the lips. The model assumes that source and filter are independent of each other. Although recent findings show some interaction between the vocal tract and a glottal source, the theory of speech production of Fant (1970) is still used as a framework for the description of the human voice, especially as far as the articulation of vowels is concerned.

The source-filter theory of speech production is exemplified by Fant representing the cavities of the mouth and pharynx by an electrical network. Current in the electrical circuit corresponds to a volume velocity in the corresponding acoustic system. This volume velocity is the product of particle velocity multiplied by the cross-sectional area of the system perpendicular to the direction of the airflow or oscillation. Voltage corresponds to pressure. The ratio of the pressure to volume velocity in terms of frequency transforms is the analogous acoustic impedance $Z = P(f)U(f)$. Output current represents the volume velocity output through the lips. Using S for source and T for the transfer function of the vocal tract filter, we can write $P = ST$ representing a speech sound. T is of course dependent on the positions of the articulators. Fant states that there is some degree of correspondence between the phonetic term phonation and the network term source and similarly between articulation and filter. This analogy implies that phonation is held apart from articulation in the sense of the generation of sound versus the specific shaping of its phonetic quality (ibid), as also mentioned in section 2.2.3.

A characteristic property of voiced speech is it's *fundamental frequency*, F_0. It is the basic property of the vocal cord sound source due to it's periodicity. Often the term pitch is used, but pitch is the perceived tonal sensation and not a property of the sound stimulus as such. However, in this work the term pitch and F_0 are often intermingled. Another characteristic is the *spectrum envelope*, which is a specification of the amplitudes of the source harmonics as a function of their frequency. This spectrum envelope reflects personal characteristics of a speaker. It also varies with voice intensity and, for instance, emotional state of the speaker (section 2.1).

It is important to recognize the existence of the different possible sources: the voice source, noise sources, and sources due to nonlinear influences (turbulence, vortices, e.g. chapter 5). In our model it is possible to turn on sources as needed.

Noise source

A noise source is a model for the acoustic disturbances within the vocal tract. In literature, it is used for generating whispering, aspirated, fricated, and exploded sounds. The source is continuous if the sound is sustainable, or it is interrupted if the shortness of the duration and the particular speed of the onset and decay are crucial.

Most noise sources are *turbulent*. Two types of turbulent noise sounds should be considered. One is the *fricative noise* produced under conditions of a relatively narrow constriction, in which case it is essentially the cavities and parts of the vocal tract in front and at the place of constriction that participate in the shaping of the

sound. The other type of sound is what could be called *open aspiration*. It is produced with greater articulatory opening than members of the class of fricative sounds. The larger opening and the occurrence of more than one source in the vocal tract, e.g., an additional glottal noise source, contribute toward emphasizing the formants that depend on the entire vocal tract and not merely those of the front parts. Frication and aspiration may occur simultaneously or in succession, or only one of the two noise categories may be present.

Voice source

Since it is probably the most important part of the vocoder, modeling of the voice source as employed in our implementation is discussed in chapter 5. Modeling a *voiced sound* in terms of source and filter will be discussed next.

Filter stage

It is possible to define the voice source by the pulsating airflow through the glottis. (The glottis represents a high impedance termination of the vocal tract and thus pressure is the dimension to describe it in.) This pulsating airflow is, in it's simplest description, a sawtooth-shaped periodic time function and can be Fourier transformed to a harmonic spectrum. The process of synthesis is determined by multiplying the filter function $T(f)$ by the harmonic source spectrum $S(f)$. When taken into account –as it is not in our implementation– the lip radiation function, $R(f)$, the resulting output spectrum of a vowel is

$$|P(f)| = |S(f)||T(f)||R(f)|. \tag{4.1}$$

This process is depicted in Figure 4.1. The phase of each harmonic is the sum of the phase of the corresponding source harmonic and the phase of the filter function. It is neglected since it does not add any substantial information [16].

The spectral peaks of the output spectrum $|P(f)|$ are called *formants* (section 2.2.4, page 14). This property results from a relatively effective transmission through the vocal tract and this $T(f)$ is assumed to be independent of the source. The location of a maximum in $|T(f)|$, which is called the *resonance frequency*, is very close to the corresponding maximum in the spectrum $|P(f)|$ of the uttered vowel. Fant notices that these should be held apart, conceptually, but for technical applications dealing with voiced sounds it is profitable to define formant frequencies as a property of $|T(f)|$. It is important, however, to be aware of the weight, so to speak, the spectral shape of $S(f)$ puts on the positioning of the formants.

As noted before, we assume that for voiced sounds the filter function is independent of the source. This implies that a formant peak will only by chance coincide with the frequency of a harmonic. The formant frequency only changes as a result of an articulatory change affecting the dimensions of the various parts of the vocal tract cavities and thus the filter function. Now, if we keep the formant frequencies constant and we double F_0, the result will be that the distance between adjacent

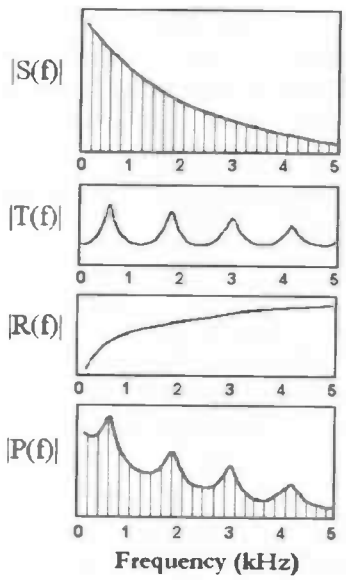


Figure 4.1: *Source-filter decomposition of the spectrum of a vowel. From top to bottom subfigure: $|S(f)|$, the source spectrum, $|T(f)|$, the vocal tract transfer or filter function, $|R(f)|$, the lip radiation function (omitted in this work), and $|P(f)|$, the resulting output spectrum of the vowel. All amplitudes are in arbitrary units.*

harmonics in the spectrum will also be doubled, and consequently the number of harmonics up to some frequency will be halved. This means that if a specific formant comes close to, for instance, the fourth harmonic at the lower pitch, it will be the second harmonic that comes closest to the same formant in the case of the higher pitch. This illustrates that formant frequency and harmonic number are different concepts and should not be confused.

Fant (1970) shows that all parts of the vocal tract contribute to the determination of all formants but with varying degrees depending on the actual configuration.

4.1.1. COROLLARY. *The intensity variations of a single harmonic or of a group of harmonics at a certain place within the frequency range depend both on the source and on the filter.*

The influence on the sound spectrum of the type of voice and the relative voice effort due to the source spectrum variations should also be considered; a reduction of voice effort, with a fixed location of all formants, leads to a decrease of the level of harmonics which is more prominent in the higher frequencies than in the lower part of the spectrum (ibid). This is due to the more steeply falling slope of the source spectrum envelope normally accompanying the lowering of the voice level. Higher formants, labeled as $F_1, F_2, \ldots$, are primarily of importance in front vowels [16]. Distances between formants in frequency average, roughly, 1000 Hz for males. This statistical average is physiologically correlated with the total length of the vocal

tract. The spread of formant data may be specifically large if all possible contextual variants of a phoneme as well as all possible speaker categories are taken into account [60]. However, given a particular context it is to be expected that a speaker will produce phonemically different sounds by means of consistent distinctions in the formant pattern (*theory of distinctive features*, ibid).

As a last remark it is to be noted that for higher frequency formants (woman, children) measuring them becomes more difficult. Looking at the tools available to estimate formants this problem becomes apparent and often higher formants are not even supplied. We find it interesting to investigate whether nonlinear techniques, e.g. the use of the Teager energy operator (section 2.5), make more robust formant estimation possible.

4.2 Formant speech synthesis

This section describes and discusses the implementation of the *formant speech synthesis* technique. Formant synthesis of speech employs the source-filter model discussed earlier. The Klatt speech synthesizer [5, 21, 26, 52] is a well-known example. One less desirable feature of the Klatt synthesizer is that it uses a rather large amount of synthesis parameters, e.g. [21]. As an extension Stevens (2002) derived a set of mapping relations, which map a smaller set higher-level quasi-articulatory parameters onto the large, lower-level set of acoustic parameters. Here we implement the original concept since we want to be free to combine parameters as needed.

The source-filter theory describes speech production as a two stage process involving the generation of a sound source, with its own spectral shape and spectral fine structure, which is subsequently shaped or filtered by the resonant properties of the vocal tract. Figure 4.2 shows our version of the model, implemented using Matlab and Simulink.

Most of the filtering of a source spectrum is carried out by that part of the vocal tract anterior to the sound source. In the case of a glottal source, the filter is the entire supra-glottal vocal tract. The vocal tract filter always includes some part of the oral cavity and can also, optionally, include the nasal cavity (depending upon whether the velum is open or closed).

Sound sources can be either periodic or aperiodic. Glottal sound sources can be periodic (voiced), aperiodic (whisper and /h/) or mixed (e.g. breathy voice). Supra-glottal sound sources that are used contrastively in speech are aperiodic (i.e. random noise) although some trill sounds can resemble periodic sources to some extent. A voiced glottal source has its own spectrum which includes spectral fine structure (harmonics and some noise and possibly AM/FM-modulations) and a characteristic spectral slope (sloping downwards at approximately -12dB/octave, for normal speech). An aperiodic source (glottal or supra-glottal) has its own spectrum which

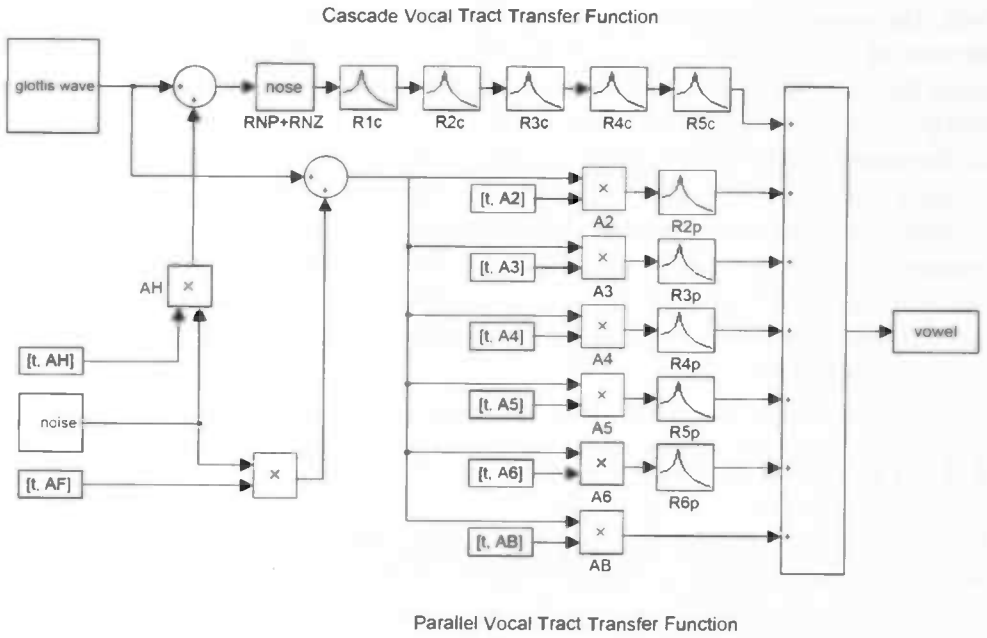


Figure 4.2: *Model of the vocoder. The filters $RNP + RNZ, R1_c, R2_c, \ldots, R5_c$ make up the cascaded vocal tract transfer function. The parallel vocal transfer function consists of the filters $R2_p, R3_p, \ldots, R6_p$. Noise can be added to both transfer functions; it will be added to the glottis wave. The parallel section needs amplification values which are absent in the cascade section. Exclusion of one of the parallel filters or noise sources is achieved by setting the corresponding amplification value to zero. RNP and RNZ constitute the nose filter. (Abbreviations: R for resonator, N for nose, P for pole and Z for zero, denoting the pole and zero resonators, subscripts c and p for cascade and parallel, respectively.) The $[t, \bullet]$ tuples refer to time related amplification values. The glottis wave is discussed in more detail in chapter 5.*

includes spectral fine structure (random spectral components) and a characteristic spectral slope. Periodic and aperiodic sources can be generated simultaneously to produce mixed voiced and aperiodic speech typical of sounds such as voiced fricatives. In voiced speech the fundamental frequency (fundamental to the perception of pitch) is a characteristic of the glottal source acoustics whilst features such as vowel formants are characteristics of the vocal tract filter (resonances).

The formant synthesizer permits the synthesis of sonorants by either a cascade or parallel connection of digital resonators, but frictation spectra must be synthesized by a set of resonators connected in parallel [26]. In his article Klatt describes a control program with which one can define acoustic parameters as a function of time. It is possible to specify formant frequencies as a set of time and value tuples. We adopted this approach here. The advantage of the parallel configuration is that the relative amplitudes of the formant peaks for vowels come out just right (ibid).

No individual formant amplitude control is needed *in theory*. When one wants to create fricatives and plosives the parallel configuration is needed too. Reason for this is the fact that it is not possible to model those sounds adequately using the relatively small amount of cascaded resonators, when the sound source is actually above the larynx! (ibid.) As a second advantage of the cascaded configuration, Klatt states that it models the vocal tract transfer function more accurately during non-nasal sonorants production. The parallel configuration is useful for generating stimuli that violate the normal amplitude relations between formants or to generate, e.g., single-formant patterns (ibid). We found that the parallel configuration was needed to synthesize vowels with a somewhat more realistic spectral envelope (appreciated on perception). To achieve this formants at a relative high frequency were introduced.

Next implementation issues are discussed in more detail. Following Klatt a cascade-parallel vocoder is implemented, as depicted in Figure 4.2. It is to be noted that among the left-out details there is the radiation characteristic of the lips.

Using Simulink

Following the Rapid prototyping approach we implemented the vocoder algorithms using Simulink®. Simulink is a software package running inside Matlab® for modeling, simulating, and analyzing dynamic systems. It supports linear and nonlinear systems, modeled in continuous time, sampled time, or a hybrid of the two. Systems can also be multirate, i.e., have different parts that are sampled or updated at different rates. Simulink provides a graphical user interface (GUI) for building models as block diagrams, using click-and-drag mouse operations. Once a model has been defined, it is possible to simulate it. Using scopes and other display blocks, the simulation results can be viewed while the simulation is running. The simulation results can be put in the Matlab workspace for post processing and visualization. It is also possible to use the obtained Simulink model outside of the Matlab environment. Since Simulink uses C code which it compiles, thereby providing the source code, one is able to create dynamic link libraries and use them in other programs.

The building blocks of the vocoder are discussed next. The Simulink implementations are given and explained.

Parameter bus

There need to be calculated three parameters for every resonator and antiresonator. In every time frame the total set of parameters might get updated. Every (anti)resonator is defined by a filter or resonance frequency, F, and a bandwidth, BW; both are determined by the filter constants A, B, and C. So, given a vector of frequencies and a vector of bandwidths, we can perform the same calculation on them and retrieve

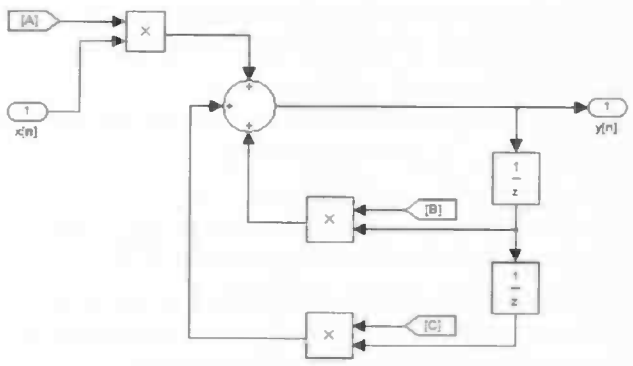


Figure 4.3: *Block diagram of a digital resonator. A, B, and C are the constants given in equations 4.3–4.5. Blocks labeled with $\frac{1}{z}$ are unit delay blocks, equivalent to the z^{-1} operator delaying the input by one sample.*

a matrix of values. The Simulink `Mux` and `Demux` blocks seem to be just perfect for organizing a lot of single wires.

Currently eleven resonator parameters can be used. So, we have to provide the model with eleven frequencies and eleven bandwidths.

Digital resonators

The main ingredients of the synthesizer are digital resonators. Figures 4.3 and 4.4(a) show the block diagram of the resonator, or bandpass filter, and it's transfer function. The output of the digital resonator is given by:

$$y[n] = Ax[n] + By[n-1] + Cy[n-2]. \tag{4.2}$$

The center or resonant frequency, F, representing the formant frequency, $F_0, F_1, \ldots$, and the bandwidth BW of the filter are determined by the constants A, B, and C as follows:

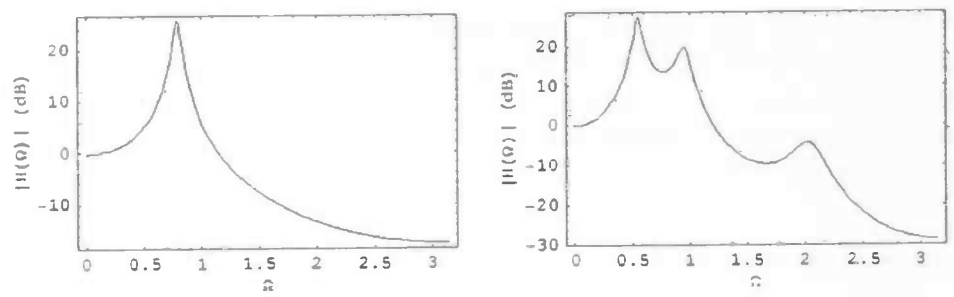
$$A = 1 - B - C \tag{4.3}$$

$$B = 2e^{-\pi BW\,T}\cos(2\pi F\,T) \tag{4.4}$$

$$C = -e^{-2\pi BW\,T}, \tag{4.5}$$

where $T = 1/F_s$ and F_s is the sampling frequency. As the smallest segment of an utterance a time step of 5 or 10 ms can be used. It is assumed that changes, at least on the filter side, occur on this time scale. Acoustic theory indicates that formant frequencies should always change slowly and continuously relative to this time interval [26]. At the start of a new segment the parameters F and BW can be altered giving new configurations for the resonators. The frequency response of the digital filter is given by

$$H(z) = \frac{A}{1 - Bz^{-1} - Cz^{-2}}. \tag{4.6}$$



(a) (b)

Figure 4.4: *(a) Transfer function of the (analog) resonator, using $F = 1000$ Hz, $BW = 100$ Hz, and $F_s = 8000$ Hz. (b) Simple example of applying equation 4.9 for three, cascaded filters.*

Next, z being equal to $e^{\jmath\Omega}$, where $\Omega = 2\pi fT$, equation 4.6 becomes:

$$\begin{aligned} H(\Omega) &= \frac{A}{1 - Be^{-\jmath\Omega} - Ce^{-2\jmath\Omega}} \\ &\equiv \frac{A}{1 - B\cos\Omega - C\cos 2\Omega + \jmath(B\sin\Omega + C\sin 2\Omega)}. \end{aligned}$$

And the gain function can thus be written as:

$$|H(\Omega)| = \frac{A}{\sqrt{(1 - B\cos\Omega - C\cos 2\Omega)^2 + (B\sin\Omega + C\sin 2\Omega)^2}}. \tag{4.7}$$

A plot is shown in Figure 4.4(a) for Ω in the range 0 to π, where we took $F = 1000$ Hz, $BW = 100$ Hz, and $F_s = 8000$ Hz. To check whether the curve is as expected we estimate the maximum gain at $\Omega = 0.8$. Now, $\Omega = \pi$ corresponds to a sinusoid with two samples per second; so $\Omega = 0.8$ corresponds to $2\pi/0.8 \approx 7.85$ samples per second. Remembering we sampled at 8000 Hz, we obtain a frequency of $8000/7.85$, or, indeed, about 1000 Hz.

Digital antiresonators

A resonator being a bandpass filter, the antiresonator must be equivalent to a band-stop filter. It's use in the original synthesizer is to shape the spectrum of the voicing source. A second use is to simulate the effects of nasalization (RNZ). The output of the antiresonator is given by:

$$y[n] = A'x[n] + B'x[n-1] + C'x[n-2] \tag{4.8}$$

where $A' = 1/A$, $B' = 1/B$, and $C' = 1/C$, using equations 4.3, 4.4, and 4.5 respectively. Note that the output signal, y, is only dependent of the input signal, x, as opposed to the equation for the resonator.

Sources

Next we need to discuss the sources. For vowels we are in need of a quasi-periodic impulse generator. If we want to generate more kinds of sounds, we need a noise

generator as well. However, we concentrate on vowels in this work and hence we will ignore this issue[1]. Note that a noise generator would be a simplification when generating fricatives. For normal voicing the source needs to generate an impulse train with frequency F_0. The number of samples between impulses, denoted as T_0, is determined by F_s/F_0.

Vocal tract transfer function

The transfer function of the vocal tract for the cascade model can be represented in the frequency domain by a product of poles and zeros. In literature it is hypothesized that five resonators are sufficient for simulating a female vocal tract. A male vocal tract, being longer on average, might need a sixth resonator. However, most literature does give us formant frequencies and bandwidths for only two or three filters. Hence, we can use these values and ignore higher formant frequencies, or we have to estimate formant frequencies and bandwidths ourselves. The latter approach is chosen for. We used a convenient speech analysis tool, `Speech Analyzer`[2], and tested recordered samples of vowels. The results were taken as a starting point and used to synthesize a set of artificial vowels. Simple perceptual testing resulted in adjusting formant frequencies and bandwidths until a set of acceptable vowels was obtained, in the sense of vowels sounding realistic.

In future work we like to design an automatic procedure to determine these vowel parameters. Nonlinear techniques, as discussed in section 2.4.2, might prove useful. It is to be noted that our set of vowels, of which table 6.1 summarizes the values used in our experiments, adapted center and bandwidths frequencies found in Dutch literature. Firstly because we used Dutch subjects in our experiments. Secondly because these values were expected and found to be reasonably sound. Looking at table 6.1 it shows that the higher formant frequencies do not differ very much. Perceptually these values were appropriate, but again, more research has to be conducted to learn how these values are related to the different vowels. As discussed in section 2.2.4, the first two or three formant frequencies and bandwidth frequencies are probably most significantly related in differentiating vowels. Higher formants contribute to more realistic vowels and introduce individual qualities.

The cascade model consisting of resonators only can be represented in the frequency domain by the product of equations for the single resonators (i.e. equation 4.6), giving, for five formants:

$$T(z) = \prod_{n=1}^{5} \frac{A_n}{1 - B_n z^{-1} - C_n z^{-2}} \tag{4.9}$$

[1]Note that noise could be used as a first, and probably over-simplified, approach in modeling vortex sound.

[2]We used Speech Analyzer version 2.7, a freeware program supplied by SIL International, Dallas.

A very simple example is depicted in figure 4.4(b). Here only three frequencies and bandwidths are defined ($\langle 700, 80 \rangle$ Hz, $\langle 1220, 140 \rangle$ Hz, and $\langle 2600, 300 \rangle$ Hz), but it shows what is possible.

Chapter 5

Glottis modeling

Difficulties increase the nearer we approach the goal.

Johann Wolfgang von Goethe 1749–1832

One of the most critical parts of the vocoder implies the modeling of the pulse train, or glottis wave, generated by virtue of the vocal folds. In literature, at least the not so recent literature, often a sawtooth-like signal was fed to the filter model of the vocal tract. However, this approach is not very convenient. Filtering is needed to influence the spectral tilt and no relation between the glottis signal and glottis parameters is possible. Glottis parameters are already mentioned in section 2.2.2, and the closed-open ratio of the glottis was demonstrated to be of significant importance concerning the spectral content of the glottis wave, e.g. [4, 9, 44, 56, 57, 59].

According to Asogawa and Akamatsu (1999) theories, previous to their work, have assumed that the shape of the glottal pulse train was a vowel-independent triangular wave. They found that literature came up with two factors differentiating voice types:

- the general spectral slope, i.e. tilt, and
- the relationship between the intensity of the fundamental frequency and its harmonics.

Motivated by Flanagan (1957) a two-pole model was expected to approximate the glottal volume flow: $U_g(z) = \frac{K}{(1-z_a z^{-1})(1-z_b z^{-1})}$. A two-pole model produces a spectral slope of -12 dB/octave. Adding more poles will 'add' -6 dB/octave for every pole[1]. Asogawa and Akamatsu provided an analytical expression for the original acoustical waveform generated at the glottis by carefully analyzing the interaction of the glottal muscle and airflow through it. Hereto they considered the kinetic energy of the glottis and the pressure of airflow from the lungs, adopting a specific theory from fluid dynamics (aerodynamics) for describing the phonation mechanism (ibid). The hypothesis is that

sound is caused by vortices due to variations in flow rate over time.

[1]Glottal pulses must be of finite duration. Therefore, an exact model of it would be a finite impulse response filter. Hence, it would have to contain only zeros, in contrast with the relation given.

Childers and Lee (1991) used the Liljencrants-Fant (LF) model and added random noise. As an advantage of the LF model they name that parameters for it can be derived by using inverse filtering. (This, nevertheless, is not a trivial matter.) They showed that glottal pulse width, glottal pulse skewness, the abruptness of glottal closure, and the turbulent noise component were important in characterizing different types of voice. Of course, equally important is to know which types of voice production they were looking at: modal, falsetto, vocal fry[2] and breathy (irregular vibration of normal vocal folds).

In this work we use the LF model to create the glottal pulses. Veldhuis (1998a) presented an alternative for the LF model. It uses the same parameters as the LF model, while the real advantage is that it is computationally more efficient. Veldhuis concluded that his model, the Rosenberg++ model (R++) since it was derived from the Rosenberg model [44], was equivalent to the LF model in practical applications. The main reason we did not follow his approach is that the computational advantage was not important for us in this work. But, we also did not want to make concessions and wanted to be able to compare results to results obtained with the same model in experiments comparable to our own as found and discussed in literature. From a perceptual point of view we probably would have had an equivalent experiment when the R++ model was adhered; Veldhuis judges it very unlikely that in the practical situation of speech synthesis small audible differences would actually yield a perceptually better approximation of real speech. This because all these models still are simple models of a complex waveform and the differences between models and waveforms are much larger than the differences between the models.

Van Dinther (2003) investigates perceptual aspects of voice-source parameters, using the LF model. He points out the the problem of *labeling* different voice qualities. One of the problems is that researchers often use the same label for different voice qualities. Another problem is that the categories are not absolute (ibid). This, in our opinion, shows that research has not yet come up with conclusive cues for most of the voice qualities. Most cues are part of a larger set of voice qualities.

We expect –and this may be a bit of a sidestep here– that

- there is a need to consider both nonlinear as linear aspects of speech,
- there is a need to come up with more (or better) criteria as to differentiate between spectral qualities or quantities,
- there is a need to consider higher level processing too.

The first point is enlightened some more in section 2.4. As an example of the second point we could mention the ongoing quest to devise a more reliable formant estimation algorithm, or maybe a more robust pitch tracker. The last remark is associated

[2]Vocal fry, or creaky voice, is the low clicking sound that vocal chords produce when pushed below their natural limit. It is effectively a toneless "rattle", rasp or roughness produced by the vocal cords at the lower end of the range which is often used as an effect in rock singing.

with the fact that humans have a lifetime of experience in scanning their environment acoustically. We combine evidence and are able to perform very well in, for instance, a cocktail party environment [3]. We nevertheless have to realize that humans are often able to classify an acoustic event even in the first fraction of its occurrence! An explanation would be that on a low level we differentiate between very basic cues, like onsets of a signal, following that we hypothesize about what event to expect, and subsequently we prune our expectation tree as more proof is gathered when the event is entering our hearing system. So, there would be a constant feedback from low-level to high-level processing, and vice versa. Another aspect that enters our imagination, so to speak, is that it is hard to be certain about when what becomes conscious to us. Like the phenomena of a *deja vu*, it might be the case that unconscious (pre)processing already took place before we became aware of the event. That is, there might have been more going on and still our experienced perception would be that we were able to classify the event practically at once. We are not aware of any literature on this subject –because we were not looking for it– but it would be interesting to set up experiments to address this issue.

The next section explains the Liljencrants-Fant model and the definitions of it's parameters. The relation between the T and R parameters is outlined, since the R parameters may form a starting point for the selection of the glottal parameters corresponding to those of natural voices [15, 16, 56, 58]. These parameters are of importance since they influence spectral tilt, an important parameter of voice quality.

5.1 The Liljencrants-Fant model

Following Schutte (1999), who gave a fine introduction on the physiology of the voice, e.g. section 2.2.2, we are interested in being able to adjust the set of T parameters: t_p, t_e, t_a, and t_0. These parameters define the opening, closing, and closed phases recognized in the glottal pulse. Figure 5.1 shows the typical representation found in literature, e.g. [16, 25, 56, 58], forming the basis of the Liljencrants-Fant (LF) model.

Obviously, the length of one glottal pulse, t_0, depends on the fundamental frequency, F_0: $t_0 = \frac{1}{F_0}$. Maximum airflow is reached at t_p with an amplitude U_0 (the actual value of it is not important for our purposes). At t_e the vocal folds collide, at which point in time the maximum excitation occurs, with an time derivative amplitude E_e. The short time interval following t_e is called the *return phase*, t_a. The interval before t_e is called the *open phase* and the interval between t_0 and $t_e + t_a$ is called the *closed phase*. During the closed phase the airflow reaches it's minimum again. In determining the mathematical relations it is often assumed that this minimal airflow equals zero, i.e. no leakage, and thus $g(t_0) = g(0) = 0$. For completeness we mention that $t_a = \frac{E_e}{\dot{g}(t_e)}$.

The return phase is modeled using an exponential decay. The time-derivative of

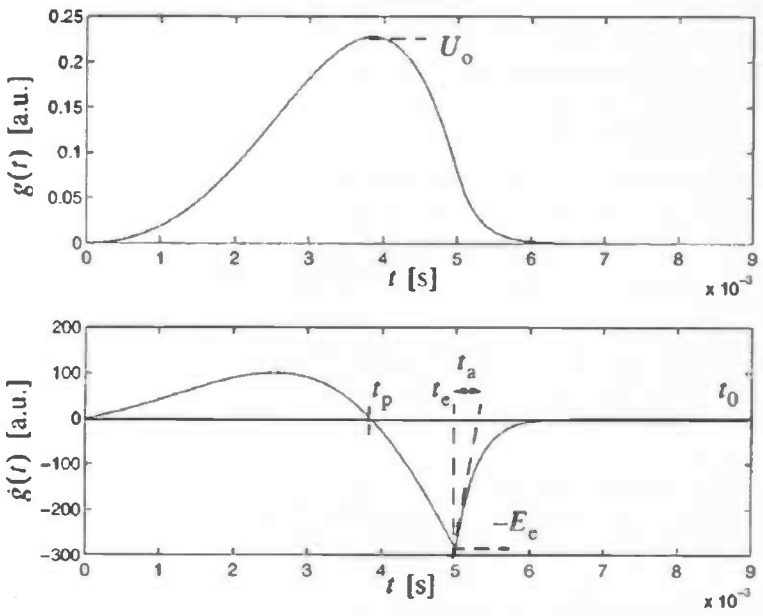


Figure 5.1: *Typical representation of the glottis pulse (top), $g(t)$, and it's time derivative (bottom), $\dot{g}(t)$. Amplitudes are in arbitrary units. The parameters t_p, t_e, t_a, t_0 and U_0 and E_e are explained in the text.*

the glottis pulse can than be expressed as:

$$\dot{g} = \begin{cases} f(t), & \text{for } 0 \leq t < t_e \\ f(t_e)\frac{e^{-(t-t_e)/t_a} - e^{-(t_o-t_e)/t_a}}{1-e^{-(t_o-t_e)/t_a}} & \text{for } t_e \leq t < t_0 \end{cases} \tag{5.1}$$

For $f(t)$ the following expression is used:

$$f(t) = B\sin(\pi\frac{t}{t_p})e^{\alpha t},$$

with B the amplitude of the glottal-pulse time derivative [56, 58]. Parameter α can be solved numerically from the continuity equation, requiring $g(0) = g(t_0) = 0$ and $g(t) \geq 0$:

$$\int_0^{t_e} f(\tau)d\tau + t_a f(t_e)\left(1 - \frac{(t_o - t_e)/t_a}{e^{(t_0-t_e)/t_a} - 1}\right) = 0. \tag{5.2}$$

Solving this equation is computationally heavy, reason for Veldhuis to come up with his R++ model. We found that, for our purposes, directly solving equation 5.2 was convenient. Solving this equation using, for example, Mathematica® requires to supply a hinting interval. In Matlab this is not needed.

5.2 Generating a glottal pulse train

Two equivalent approaches to generate the glottis wave were tried. Using Simulink there is the advantage of being able to add tools like oscilloscopes, periodograms 'on the fly'. It is relatively easy to understand the whole picture and you get some of your presentation sheets for free. However, using Matlab is advantageous when speed issues play a role. After having tried numerous methods to generate a flexible glottis wave, we wrote Matlab (.m) code to do the job. One very interesting method will be discussed briefly in the next section. This method implies using a shaping function, making it possible to change period, amplitude and spectral centroid in real-time, without the need for recalculating various parameters all the time. In the end we abandoned this approach since it was not possible to compute a sufficient amount of harmonics; matrix operations became erroneous making it impossible to obtain correct pulse shapes.

We synthesized a glottis wave as follows. First we choose a set of R parameters. These were transformed into T parameters using the relations[3]

$$\begin{aligned} t_e &= r_0\, t_0, \\ t_a &= r_a\, t_0, \\ t_p &= \frac{t_e}{1 + r_k}. \end{aligned}$$

Next α was computed using the Matlab function `solve`. Then a glottis pulse could be build by integrating equation 5.1. As integration function we used Matlab function `cumtrapz`. This method works fine when a sufficiently high sampling rate is used, that is, $\dot{g}(t)$ has to be defined in high detail around its discontinuity to let integration come up with a correctly shaped glottis pulse. We used a sampling rate of 6 MHz. Of course, resampling is needed to end up with a convenient sampling rate for experimenting, e.g. 44.1 kHz. After resampling the glottal pulse is removed from it's DC component by subtracting the mean value of the pulse. Without this step a zero frequency will appear in our synthesized vowels. Finally, glottis pulses were concatenated, as needed, to form a pulse train.

In our experiments we did not need pulse trains consisting of different shaped pulses (in a single vowel). In that case we would have needed to calculate every pulse using another set of r_o, r_a and r_k, and it is likely that it would be profitable to use an other synthesizing scheme.

We are left with the problem of finding R parameters corresponding to those of natural vowels. In Veldhuis (1998a) a so-called shape parameter, r_d, was used. Together with the fact that there exist simple statistical relationships between the shape parameter and the R parameters, convenient parameters can be chosen. You

[3] OQ $\equiv r_0 = t_e/t_0$ is the open quotient, a term frequently used in literature. It denotes the fraction of the period the glottis is open. Other parameters often found in literature are the closing quotient, CQ, and the speed quotient, SQ. CQ reflects the transient character of the glottis pulse and SQ reflects the asymmetry of the glottis pulse.

are left with the determination of the shape parameter, however. In this work we borrowed R parameters as found in literature, e.g. [56]. Also preliminary perceptual testing of R parameters was used. Chapter 6 gives more details about our choice of parameters.

The next section discusses the shaping function and the problems we encountered.

5.3 Shaping a template pulse

This section is based on the work of Schoentgen, [46, 47]. We briefly discuss the shaping function and show the problem we encountered. We finally abandoned this very interesting shaping method instead of trying to solve the problem, because of a lack of time.

Schoentgen gives the following abstract: "A shaping function model is a non-linear memoryless input-output characteristic that transforms a simple harmonic into the desired output. The model can be fitted linearly to observed or simulated template cycles. The instantaneous values of the excitation cycle centroid, amplitude as well as length, and the cues for phonatory identity are set via distinct parameters. The synthetic phonatory excitation signal is zero on average, as well as identically zero when the glottal airflow rate is constant." This is what we really want!

The model for the glottal wave, s_g, is based on a power series:

$$s_g(n) = c_0 + c_1 x(n) + c_2 x^2(n) + \ldots + c_M x^M(n),$$

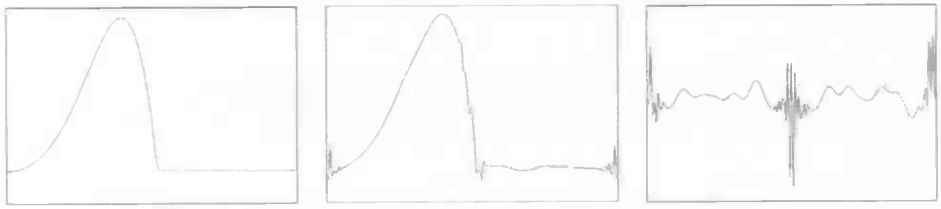
with M the order of and c_i the coefficients of the power series, and n the time index. Using a set of considerations (consult article) Schoentgen comes up with the next model:

$$s(n) = G\left\{\sum_{i=0}^{M} c_i A \cos^i([\theta(n)]) + A\sin[\theta(n)]\sum_{i=0}^{M} d_i A^i \cos^i[\theta(n)]\right\}. \tag{5.3}$$

The instantaneous frequency of an output cycle is the instantaneous frequency of the driving cosine, $\theta(\cdot)$ (ibid). Amplitude is scaled by G and cycle shape depends on A and the coefficients c_i and d_i. The latter coefficients are computed by inverse matrix operations, based on the following relations:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \ldots \\ a_M \end{pmatrix} = M_e \begin{pmatrix} c_0 \\ c_1/2 \\ c_2/4 \\ \ldots \\ c_M/2^M \end{pmatrix} \text{ and } \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \ldots \\ b_{M+1} \end{pmatrix} = M_o \begin{pmatrix} d_0/2 \\ d_1/4 \\ d_2/8 \\ \ldots \\ d_M/2^{M+1} \end{pmatrix}. \tag{5.4}$$

The values of a_i en b_i are Fourier coefficients, obtained by Fourier transforming a template glottis pulse. M_e and M_o are defined with the use of the Pascal arithmetical triangle (details in Schoentgen). Since one wants to generate glottis waves with a sufficiently high number of harmonics a problem is encountered when trying to solve equations 5.4, applying the inverse matrices M_e^{-1} and M_o^{-1}. When we use a sampling frequency of, say, 16 KHz, we would like to find harmonics of the fundamental frequency up to 8 KHz. Using a fundamental frequency of 100 Hz, this boils down to the use of $M = 80$ Fourier coefficients and 81×81 matrices. We found that 40 harmonics were about the limit. Obviously, the frequency spectrum of the synthetic glottis wave stops abruptly above $M\times$ fundamental frequency. We end this section by showing three glottis pulses, synthesized using the shape function for $M = 40, M = 50$ and $M = 60$, plotted in Figure 5.2.



$M = 40$ $M = 50$ $M = 60$

Figure 5.2: *Glottis pulse synthesized using the shape function, for $M = 40, M = 50$ and $M = 60$ Fourier coefficients, from left to right. Clearly we arrived at a limit.*

5.4 Spectrum centroid related to r_o, r_a and r_k

To obtain some measure of the effect of altering the R parameters, r_o, r_a, and r_k, the spectral centroid of the glottal pulse is estimated. Changing one of the parameters may change the spectral tilt, which may result in a more richer account of harmonics in the glottal pulse train, and moving the spectral centroid to a higher frequency. The centroid is obtained by calculating the statistical first moment of the signal spectrum. There are of course several ways to do this; the method used here is by calculating and correctly scaling the output of the FFT function to obtain a meaningful power versus frequency plot[4] and use the power and frequency values to do the statistics.

To test the relation between the R parameters and the spectral content of the glottal pulse we generated 216 different pulses. The R parameters were all changed in six steps, the r_a and r_k parameter in equally sized steps, but the r_o parameter was divided by two every step. This was done because a preliminary test showed that the r_o parameter statistically proved to have the most impact on the value of the centroid[5]. Figure 5.3 shows the results. It is compiled of 36 sub figures of which

[4]By Atsushi Marui and based on the Matlab Technical Note 1702 (published at http://www.mathworks.com/support/tech-notes/1700/1702.shtml).

[5]Data analysis is done using R, which is a language and environment for statistical computing and

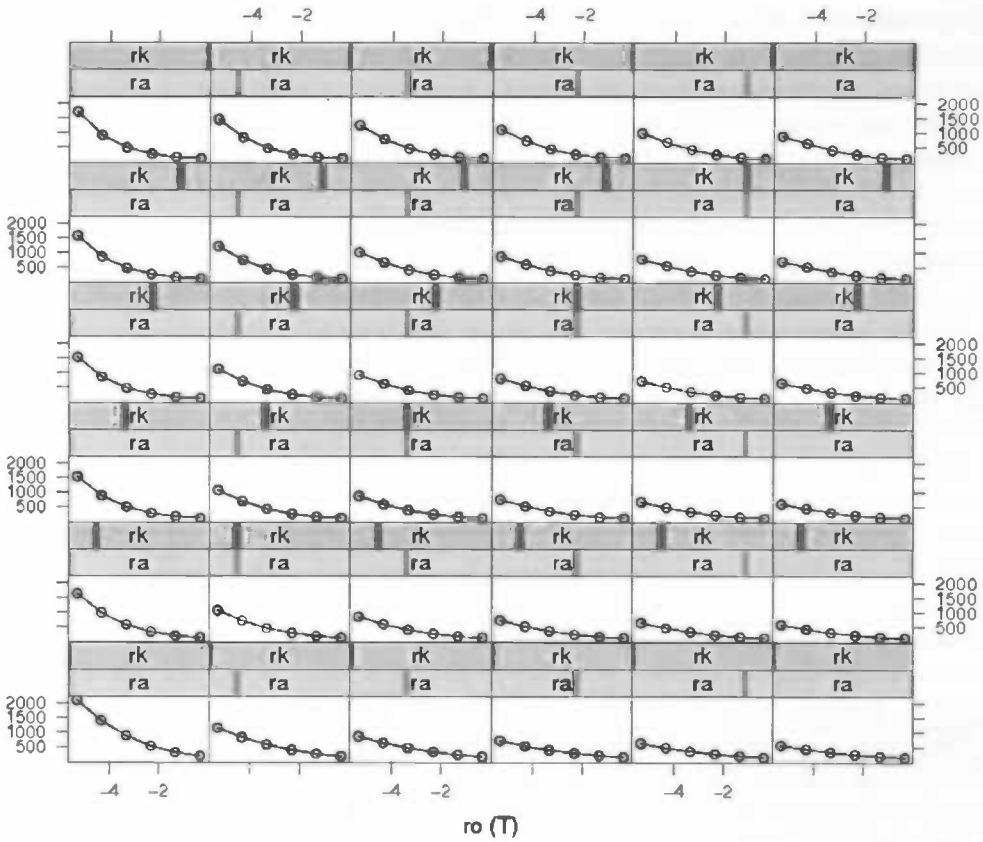


Figure 5.3: *Effect of r_o, r_a, and r_k on spectral centroid. Horizontally the $\log_2$ of the r_o parameter ($r_o \in \{0.025, 0.050, 0.100, 0.200, 0.400, 0.800\}$) and vertically the calculated centroid is given. The applied r_a and r_k are given on top each subfigure: $r_a \in \{0.0030, 0.0104, 0.0178, 0.0252, 0.0326, 0.0400\}$ and $r_k \in \{0.10, 0.26, 0.42, 0.58, 0.74, 0.90\}$. Read from left to right r_a increases, as does r_k from bottom to top.*

the horizontal axis shows the $\log_2$ of the r_o parameter ($0.025 \leq r_o \leq 0.8$) in relative duration of the period, T, of the glottal pulse. The vertical axes show the calculated centroid (Hz). The applied r_a and r_k are given above each figure: $0.003 \leq r_a \leq 0.04$ and $0.1 \leq r_k \leq 0.9$. Read from left to right r_a increases, as does r_k from bottom to top.

The figure clearly shows that shorter open phases (r_o) always increase the spectral centroid. The question is whether very small values sound natural as the glottal pulse becomes very steep at that point. Figure 5.4 shows the relation of the R parameters on pulse shape. Increasing r_k, or skewness, lowers spectral centroid. This effect is more apparent for smaller values of r_o though. Decreasing the return phase,

graphics. See for instance Baron and Li (2004).

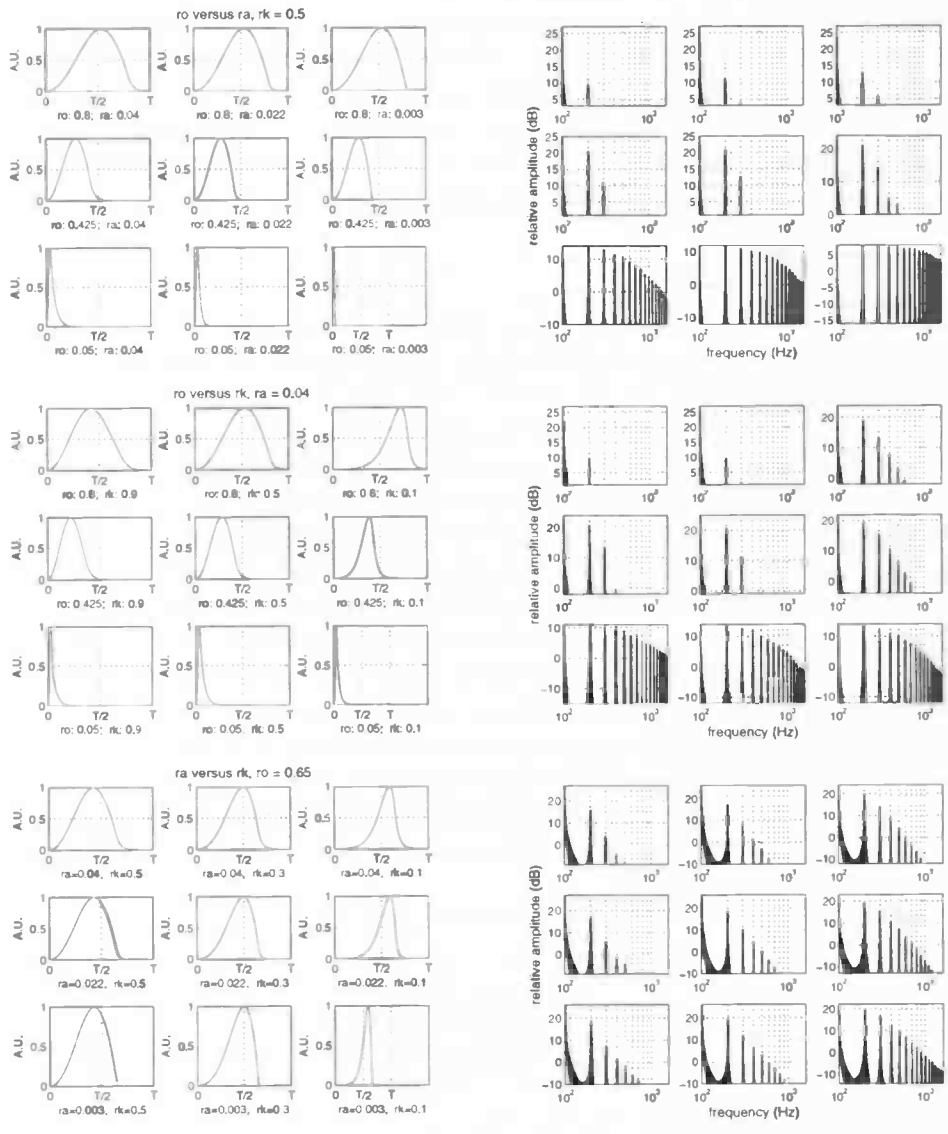


Figure 5.4: *Effect of r_o, r_a, and r_k on normalized glottal pulse shape (a) and spectral tilt (b). On the horizontal axes the first parameter is held constant, on the vertical axes the second parameter is held constant. Every sub figure holds the third parameter constant. Note that for the spectral tilt figures, i.e. column (b), the parameter values are not repeated; however, the left and right corresponding sub figures can easily be mapped. Further, the spectral figures contain plots of four octaves, i.e. 100...1600 Hz, and the y-axes are scaled such that comparison of the tilts is made possible.*

r_a, also significantly increases the spectral centroid.

At first, it is surprising that Veldhuis (1998b) concludes that r_a has the highest

spectral significance, that the spectral significance of r_o increases with increasing r_a and that the spectral significance of r_k remains at a constant low level. But, taking a closer look, e.g. at Figure 5.3, reveals the emphasis of r_a on the spectral centroid. At any rate, fact is that both r_a and r_g have a great impact on the change of the glottal airflow, whilst r_k only plays a minor role. Hence, when synthesizing vowels for our experiment, we should primarily change the values of these two parameters.

Labeling voice qualities

In literature, a large number of labels for different voice qualities have been proposed. Confusingly, researchers often use the same label for different voice qualities, and different labels for one and the same type of voice [56]. Another problem in labeling voice qualities is the fact that the categories are not absolute (ibid). Voice quality refers to characteristics such as whisperiness, harshness, nasality, pitch, and loudness. Keller (2005) examines articulatory and acoustic correlates of voice quality in terms of an initial unified scheme, which was proposed by Laver and summarized in his article. For example, the voice qualities whispery, creaky, breathy, harsh, tense and lax are voice qualities which are mainly influenced by the larynx of the vocal system. In speech synthesis, control of voice quality will improve the naturalness of synthetic speech. In this work, referring to chapter 6, we only use four labels which are expected not to get confused. These labels are `neutral`, `fear`, `cold anger` and `hot anger`. For our experiment Dutch subjects were chosen and thus Dutch labels were used. These Dutch labels were even more unambiguous and sufficed for our purposes. When subjects have to differentiate between more emotions, labeling voice quality becomes more of a problem and certainly would need more attention.

Chapter 6

Experiment

A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.

Albert Einstein 1879–1955

By means of a psycho-acoustic experiment we investigated whether synthetic vowels generated with our model at various choices of parameters can be perceptually discriminated on emotional content. A few aspects were of concern. First of all isolated vowels were chosen because a psycho-acoustical comparison of isolated vowels is more critical with respect to discrimination than the comparison of synthetic speech in which other synthetic artifacts, phoneme transitions, and the context may mask the perceptual differences [9, 58]. However, a small shaping of the vowel was permitted, because otherwise vowels would sound to synthetic and it was expected that this would bias the experiment too much. The purpose of this first experiment is to determine the practical and scientifical relevance of the vocoder by means of synthesizing vowels with different intended emotional content. Hereby our main assumptions are reflected in the following hypotheses:

6.0.1. HYPOTHESIS. *Increasing arousal is reflected in the increasing spectral statistical moment of a vowel.*

6.0.2. HYPOTHESIS. *Increasing arousal may eventually result in loss of control of the speech production, which is reflected in modulations of the speech signal.*

6.1 Method

The shaping meant that the fundamental frequency of every vowel was changed by a small amount (+10% ... -10%) during it's 'utterance'. Figure 6.1 shows the shaping function used. Vowel length differed between 0.5 and 1.5 seconds, randomly. In trying to let the shaping function be of equal influence, the onset and offset for each vowel is the same and during the time between the onset and offset (depicted as the two vertical lines in the figure) pitch is kept practically constant. In experiments like the one performed by Veldhuis (1998a) segments of 0.3 seconds were synthesized. In their experiment subjects had to discriminate between vowels generated by the LF model and their R++ model (see chapter 5), in a three-interval three-alternative forced-choice paradigm. The subjects task was to indicate the segment generated

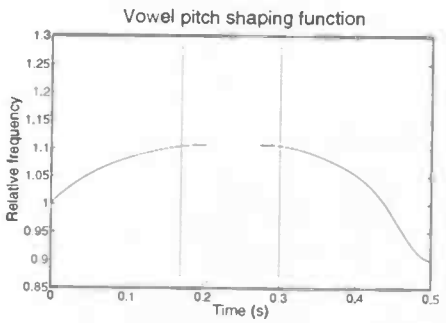


Figure 6.1: *Vowel fundamental frequency follows the depicted shaping function for more realistic sounding vowels. Vowel lengths are between 0.5 and 1.5 seconds. The vertical lines mark the region in which pitch does not change significantly. The example shows the function for a vowel of 0.5 seconds in length.*

with the LF model. Here no such approach is chosen. The purpose of this first experiment is to find out whether we are able to synthesize vowels with our vocoder with different emotional content. Indeed, one option was to let subjects compare short segments only differing on a single aspect (e.g. pitch or spectral content), but we chose the following paradigm.

We created a dataset consisting of the Dutch vowels /a/, /e/, and /u/. Formant frequencies were fixed at the values given in table 6.1. Fundamental frequency was increased in five steps from 180 Hz to 450 Hz, that is, every vowel had a fundamental frequency $\in \{180.0, 247.5, 315, 382.5, 450.0\}$. Glottal pulse shape was varied using the R parameters (chapter 5). Starting points were the values Van Dinther

Table 6.1: *Formant frequencies and bandwidths (in Hz) as used in the experiment*

vowel	F_1	F_2	F_3	F_4	F_5	F_6
/a/	770 (30)	1450 (60)	2450 (150)	3400 (200)	3700 (200)	4900 (1000)
/e/	442 (80)	2044 (100)	2601 (82)	3393 (250)	4268 (200)	4900 (1000)
/u/	385 (65)	1025 (210)	2144 (140)	3302 (250)	4218 (200)	4900 (1000)

(2003) found in his thesis. Van Dinther used labels like `lax`, `modal` and `tense`. We defined four different shapes meant to show an increase in spectral content (see section 5.4). Figure 6.2 shows plots of the four shapes we used in our experiment ($t_0 = 1/180$ s). Shapes were concatenated to form a glottal wave of the expected length. Table 6.2 shows the parameters that define the four shapes. The last column of the table contains the determined value of the spectral moment of each shape, using the method of section 5.4. As expected, spectral moment, which is related to

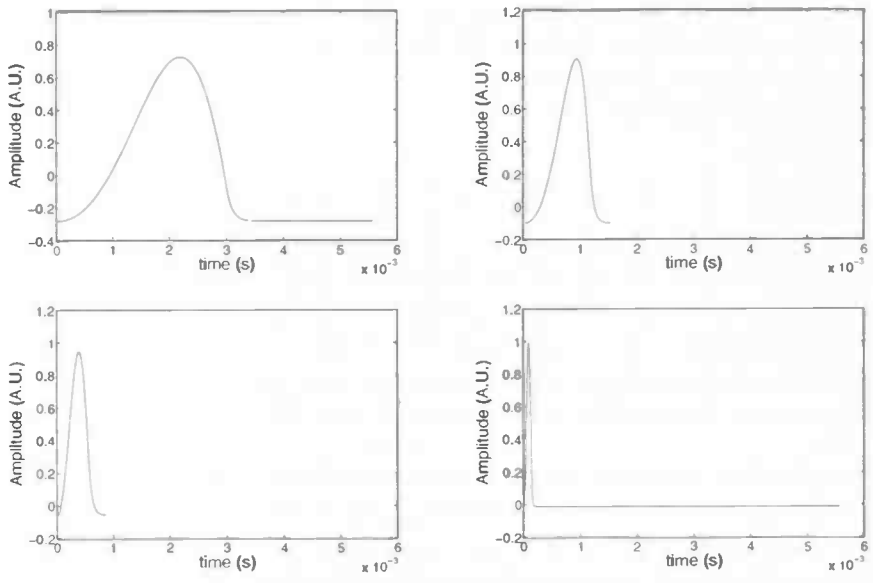


Figure 6.2: *The four shapes that formed the basis for the glottal waves used in the experiment. Fundamental frequency is 180 Hz.*

spectral tilt, increases with shape number[1]. Further we defined *jitter* and *shimmer* as cues. We expect that (most) cues can be related to source changes. A simple definition to test jitter is to let the fundamental frequency vary by some small amount, every cycle. The same goes for shimmer. In our experiment we let jitter vary by 0%, i.e. no jitter, 4%, and 8% maximal. For shimmer 0%, 15% and 30% variation was used. For a given jitter a random Gaussian distribution was generated consist-

Table 6.2: *Parameters determining the four shapes used. The last column shows the spectral moment of each shape, using a concatenation of 100 pulses.*

shape	r_o	r_a	r_k	Spectral moment (Hz)
1	0.54	0.018	0.37	229
2	0.21	0.011	0.25	483
3	0.10	0.010	0.42	757
4	0.03	0.002	0.58	2891

ing of values between -jitter and +jitter, such that mean frequency did not change. The same considerations concern shimmer. Figure 6.3 depicts how a pulse train is formed by concatenating pulses with period $t_o + \delta t_0$.

So, our data consisted of $3 \times 5 \times 4 \times 3 \times 3 = 540$ synthesized vowel samples (i.e. the number of different vowels, fundamental frequencies, pulse shapes, jitter

[1]Spectral moment was calculated before and after resampling (recall section 5.2). The values obtained were the same (±2) for all shapes.

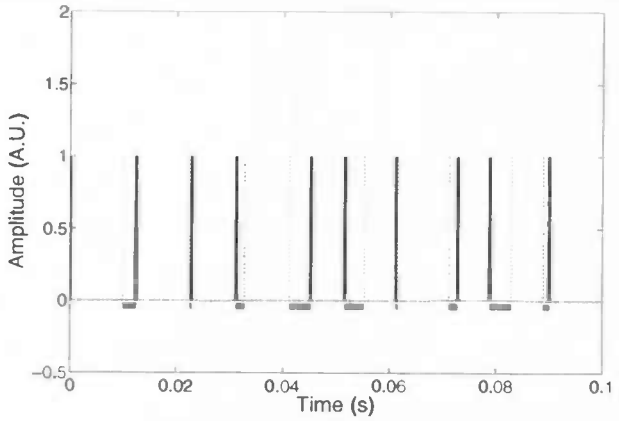


Figure 6.3: *Definition of jitter for our experiment. Vertical solid lines are the moments of a new glottis pulse cycle. The vertical dashed lines would have been the starts of new cycles when jitter would have been 0 counting from the previous cycle. The thick horizontal lines indicate the amount of jitter, being the difference of the two related vertical lines. Mean period equals 0.01 seconds; after nine cycles the last pulse starts at 0.9 seconds (for illustration).*

and shimmer percentages). The samples were randomly presented to the subjects. All subjects used the same headphone and were asked to chose a quite place to sit to do the experiment on a laptop. During the experiment subjects listened to each fragment and were asked if they could identify the fragment with a vowel $\in$ {/a/, /e/, /i/, /o/, /y/, /u/}. Then they were asked if they found that there was some emotional content $\in$ {ColdAnger, Fear, HotAnger} in the fragment. If so, they were asked to indicate the degree of perceived emotion on a scale of 1 to 4. If a fragment was identified as a vowel subjects were finally asked whether they found the fragment sounded realistic, i.e. human-like.

It should be mentioned that the listening test results should not be interpreted as determining the best choice of glottal parameters or even the best glottal model for the voice qualities we considered. Although a preliminary listening test was performed and the set of samples was found convenient, analysis of the results supplied us with valuable information giving directions for future perceptual experimenting.

A final remark: no training was allowed, therefore the data is searched for evidence of training effects. Interviewing the subjects after they had participated the listening test, gave us the idea that there was a kind of learning present, at least for part of the subjects. It seemed as if those subjects made some rules during their session and used them to decide on emotional content and realism.

The next sections show the results of our analysis of the data collected during the experiments. First of all we check how our subjects judged the vowels: did they agree on vowel identity? What was their opinion on the vowel being realistic? An interesting aspect is the fact that two subjects have hearing problems. One of

them uses hearing-aids and the other claimed he has some damage due to loud (fabric) noise. It is to be noted that the latter person does not show any hearing problems in his social behavior (the former –unfortunately– does!). Finally the data is statistically analyzed[2] testing for significance of the cues we put into the vowels.

6.2 Results

By means of factorial ANOVA we carried out univariate tests of significance for emotional content, the degree of perceived emotion, whether the perceived vowel sounded realistic or not, and whether the perceived vowel was the intended vowel. These dependent variables were tested against the predictors shape, fundamental frequency or pitch, jitter and shimmer. Categorical variables (e.g. emotion, realism, vowel) were not recoded to numericals.

Vowel perception

We intended to synthesize three Dutch vowels, /a/, /e/, and /u/, of which the formant frequencies are given in table 6.1. The following confusion matrix shows how well the subjects agreed with our intentions:

Table 6.3: *Confusion matrix showing intended vowels (vertical) versus percepted vowels (horizontal) for all participating subjects*

	/a/	/e/	/i/	/o/	/y/	/u/	?	sum
/a/	**1026**	124	0	39	173	27	51	1440
/e/	0	**357**	415	28	358	125	157	1440
/u/	35	37	10	150	**239**	**684**	285	1440
								4320

The columns give the choice subjects had in choosing which vowel they heard. They were also given the option None ('?' in the matrix) when they did not recognize any vowel at all. There is, as expected, some confusion between the percepted vowels /u/ and /y/. Adding their scores together a combined recognition of 923 is obtained. Things are bad for the intended /e/, in practically 25% of the cases it is confused with the /y/ and, even worse, it is confused with the /i/ almost 29% of the time. Only 25% of the intended vowel /e/ was percepted as such, against 47.5% for the /u/ and 71% for the /a/. When we combine the /y/ and /u/ recognition is 64%, still almost 20% is not recognized.

It would be interesting to see if (part of) the confusion is related to the hearing problems two of the subject suffer. Figure 6.4 shows density plots reflecting confu-

[2]Statistica 6.0 from StatSoft, Inc., Tulsa, USA is used for analysis. Also R is used to confirm results. R is a language and environment for statistical computing and graphics.

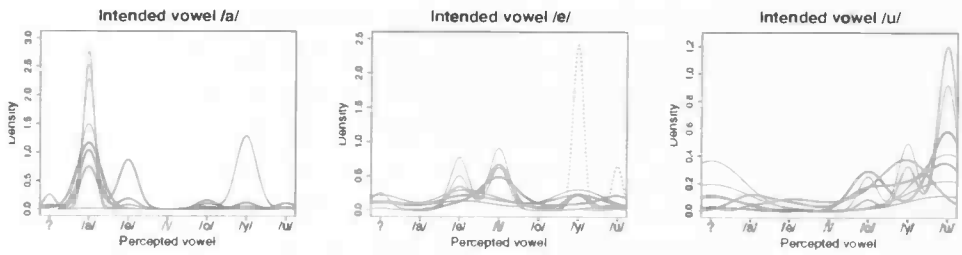


Figure 6.4: *Density plots showing the confusions for every subject for each of the three synthesized vowels. Horizontal are the perceived vowels; no vowels recognized is denoted by a '?'. On average the /a/ was perceived best.*

sion. The plots show no significant deviation for the subject with the hearing-aids (green curve). For the other subject (blue curve), with his supposed hearing deficiency, significant deviations are observable. He confuses[3] the /a/ with the /u/ and the /e/ with the /u/ and/y/. He only perceived two /a/'s and no /e/. Further, for instance, one subject (red curve) confuses the /a/ and the /e/. It seems that only one of the hearing-problem persons does influence the overall score. Looking at the data, this subject only identified 166 vowels correctly (joining the vowels /u/ and /y/), while the scores for the other subjects are 233, 246, 263, 300, 338, 377 and 383. Scores range from 31% to 71% correct.

Significance tests show that the choice of vowel (which vowel did a subject perceive, if any?) was significantly related to shape and also, but less evident, to pitch: $F(3, 4140) = 6, p = 0.00027$, and $F(2, 4140) = 2, p = 0.05$, respectively[4].

The fact that there is more confusion for the /e/ than there is for the /a/ is explained by the much higher first formant frequency for the vowel /a/. This becomes apparent inspecting table 6.1 or figure 2.7. When F_0 increases and approaches

[3]Confusion is not the correct term in some sense, since we labeled our synthesized vowels using supposedly correct formant data.

[4]These results are obtained using Factorial ANOVA under Statistica. This way higher-order interactive effects of multiple categorical independent variables (factors) can be analyzed. On occasions R has been used: first a linear mixed-effects model is fitted, using function `lme`; data was grouped on subject. Next a model is chosen, using `stepAIC`, by stepwise model selection by exact AIC. The purpose of this stepwise regression is to find the smallest set of predictors that do a decent job of predicting the dependent variable. There are a variety of criteria to use when determining if a model is doing a decent job of predicting. `stepAIC` is a stepwise procedure, implemented in R, based on Akaike's Information Criterion (AIC). It does not require a nested model, contrarily to the F-test. Stepwise regression is a controlled elimination of predictor variables, based on lowest t-statistic. Elimination is continued until all variables left in the model are significant. An example of model specification is: `vowel ~ shape + pitch + jitter + shimmer`. `steapAIC` can perform forward and backward elimination, or both. Comparing methods (forward, backward) may give you confident in the model when they give the same result. The significance of the partial regression coefficients is tested using t-tests: which variable is particulary important in terms of unique contribution to the variability of the independent variable (e.g. vowel). ANOVA results are different; here F-tests evaluate whether or not adding predictor variables significantly improve the fit of the model containing all previous predictors. Linear regression is about finding a model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, where β_k are the partial regression coefficients.

F_1 for the /e/ (or /u/ and /y/) F_1 becomes effectively eliminated, so to speak, for the latter vowels. This is not the case for the /a/. Also, higher pitch results in less harmonics, so the formant-envelope for the vowels /e/, /u/ and /y/ become more alike and less discriminable.

Realism of the vowels

In order to be able to say something about the usability of the vocoder we would like to know the opinion of the subjects when it comes to the vowels sounding realistic, or not. Only when subjects perceived a vowel they were asked to judge realism. When they did not perceive a vowel the question of realism would be an absurd one, of course. Figure 6.5 shows the results for our test group. There are two negative outliers; one of them is JB who has a hearing problem, the other, CtB, is someone who is used to work with children with hearing problems. In some sense the result of subject RS represents a positive outlier. RS claims to suffer some hearing problems too. The mean percentage of perceived realistic fragments is given

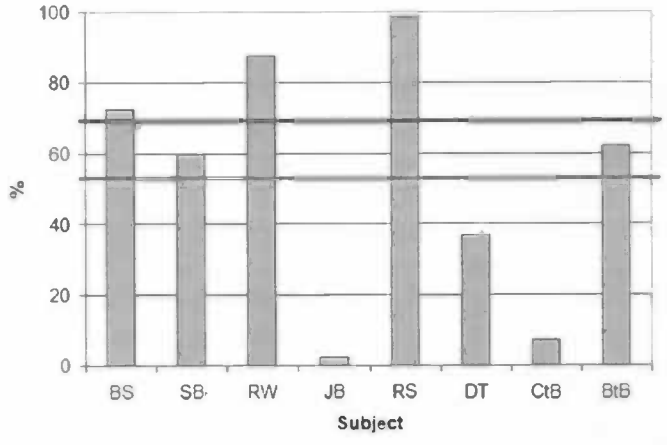


Figure 6.5: *The percentages of fragments per subject perceived as realistic. The lower, red line indicates the mean percentage for all subjects; the upper, green line shows the mean for all subjects but the two negative outliers (JB and CtB).*

by the lower, red line. Leaving out the two low scores, we obtain a mean represented by the upper, green line. The means are $53 \pm 35\%$ and $70 \pm 22\%$ respectively. To assess whether these relatively large standard deviations are representative we should use more subjects. However, the obtained values indicate that the vocoder is able to produce useful data because of the positive correlations with shape and pitch (next section). That is, for some subset of our data, depending on the parameter values used, it was increasingly likely to be perceived as realistic vowels. Thus, using correct parameter values results in high realism, and therefore the vocoder does succeed in synthesizing useful data.

Emotion, degree and realism

The next three tables (table 6.4 – 6.6) show the results for each dependent variable, including interaction effects of the predictors, except for intended vowel. The latter variable is not related to one of the effectors ($F = 0$ and $p = 1$). Subjects could chose between the Dutch labels `neutral`, `boosheid`, `angst`, and `woede`, loosely translated as `neutral`, `cold anger`, `fear`, and `hot anger`. For statistical analysis these categorical labels are used, but to assess the effects of numerical labeling

Table 6.4: *Effects on emotional content of the synthesized vowel. Emotion categories are* $\in$ {None, ColdAnger, Fear, HotAnger}. *(df = degrees of freedom, * indicates statistical significance.)*

Effect	df	F	p
pitch	4	109	< 0.000001*
shape	3	5	0.001*
error	4140		

Table 6.5: *Effect on the degree of emotion perceived in the synthesized vowel. Emotion categories are* $\in$ {None, ColdAnger, Fear, HotAnger}. *(df = degrees of freedom, * indicates statistical significance.)*

Effect	df	F	p
pitch	4	57.4	< 0.000001*
shape	3	4.3	0.005*
shimmer	2	3.5	0.03*
jitter	2	3.4	0.04*
other interactions	≥ 6	< 0.9	> 0.4
error	4140		

Table 6.6: *Effect on the amount of realism perceived in the synthesized vowel. Emotion categories are* $\in$ {None, ColdAnger, Fear, HotAnger}. *(df = degrees of freedom, * indicates statistical significance.)*

Effect	df	F	p
shape	3	11	< 0.000001*
pitch	4	3	0.02*
error	4140		

of the emotions we also did the same tests after translating the emotion categories into integers $\in \{0, 1, 2, 3\}$. One might expect that, if one had to apply an emotion scale, it might be possible to order the labels as neutral < fear < cold anger < hot anger, in increasing order of arousal But, possibly cold anger and fear should be swapped. This ordering is inspired by Feeltrace, initiated by Cowie et al. (2000) who use two dimensions to categorize emotions. Two dimensions is likely to be insufficient and the ordering of our four labels might not make much sense. To account for this, we changed the ordering of the numerical emotion labels in our data to investigate it's effect. We changed ordering for fear, cold anger and hot anger; neutral was always the first emotion. So we obtained six different orderings. For degree and realism shuffling labels had no effect. However, for emotional content ordering matters. Pitch (fundamental frequency) was almost always the best predictor; only when ordering was fear < hot anger < cold anger, then shimmer was the best predictor, $F(2, 4140) = 4.0, p = 0.02$, and pitch then was the second best, $F(4, 4140) = 2.7, p = 0.03$. However, overall significance was less than for every other ordering. The effect of shape was not significant, but for the orderings neutral < cold anger < hot anger < fear or neutral < cold anger < fear < hot anger. For the former ordering significance of the predictors shimmer and jitter decreases, for the latter this is not the case. Based on this analysis the most predictive ordering –in our experimental setup– would be neutral < cold anger < fear < hot anger. When we group the emotions (cold and hot anger and fear) then the predictors pitch, jitter and shimmer are still significant for emotion, but shape is not anymore. For degree and realism nothing changes. This shows that shape is important in the discrimination of the type of emotion.

Of course, these results cannot be generalized, but it shows that labeling emotions needs attention. When we use R to do the significance test using a linear mixed-effect model, then the effect of the predictors on realism include also the shape×shimmer interaction as significant.

If only data records are used in which vowels are not recognized ($N = 493$) then shape is not a significant predictor for emotion. Pitch is, as it is for degree. If only data records are used in which vowels áre recognized then shape stops being a significant predictor for degree, $F(3, 3647) = 0.7$, $p = 0.5$; jitter and shimmer p-values stay practically the same, however.

Apparently emotional content can be perceived without sound being recognized as a vowel (although the perceived sound clearly is tonal in nature). Leaving out the records for no emotion results in almost no vowel not being classified (about 1% instead of more than 11%). Hence, emotion is related to the fact that a vowel is recognized, which is not in contradiction with the previous observation that emotion is perceivable in non-vowels.

Emotion discrimination

To discriminate between emotions, i.e. which of the cues shape, pitch, jitter and shimmer are most decisive for which emotion, first of all interaction response tables

Table 6.7: *Frequency table showing the number of times subjects perceived emotion as a function of shape.*

	Emotion			
shape	None	Cold Anger	Fear	Hot Anger
1	639	191	160	90
2	611	193	194	82
3	613	168	204	95
4	589	161	208	122
	2452	713	766	389
fraction	57%	16%	18%	9%

have been drawn. From these plots –take a look at figure A.1, appendix A– a few trends can be observed. First of all, the number of neutral observations decreases with increasing pitch. This is true for all jitter and shimmer values. These observations confirm the statistical significance of pitch on emotion.

For the emotions cold anger and hot anger, trends are not that obvious when looking at the interaction plots. For fear, a strong positive relation is shown for all values of jitter, shimmer and shape (fig. A.2). To be complete, and because they show interesting differences between the emotions, the remaining interaction plots, for cold anger and hot anger, are given in respectively figure A.3 and figure A.4. These figures show a (light) negative trend for pitch on cold anger, and a light positive trend for pitch on hot anger.

For the other cues it is not easy to determine trends on visual inspection alone. It seems that for the dependent variable degree all positive values (since degree = 0 only happens when emotion neutral was chosen) show a light positive trend. Strikingly, this trend is more prominent for degree = 3 than for degree = 4. This might

Table 6.8: *Comparing pairs of emotions, leaving other emotions out of the data. (N: Neutral, C: Cold Anger, F: Fear, H: Hot Anger; 0 denotes a $p < 0.0001$.)*

	p for emotion pairs					
Predictors	N–C	N–F	N–H	C–F	C–H	F–H
shape		0.04	0.015	0.014	0.007	
pitch		0	0	0	0	0
jitter			0.0002		0.0005	0.02
shimmer			0		0.0008	0
observations	3165	3218	2841	1479	1102	1155

Table 6.9: *Dependent variable degree is fitted for the independent variables shape, pitch, jitter and shimmer. Horizontal are the emotions, which are kept constant for each analysis. Statistics analyzed using R. (* denotes significance)*

	p for degree per emotion					
	Cold Anger		Fear		Hot anger	
Predictor	F	p	F	p	F	p
shape	2.3	0.13	2.2	0.14	8.2	0.004*
pitch	2.5	0.11	91.0	< 0.000001*	8.0	0.005*
jitter	1.8	0.19	2.8	0.01*	20.5	< 0.00001*
shimmer	2.4	0.12	0.18	0.67	0.4	0.5
jitter:shimmer					4.4	0.04*
shape:shimmer			2.0	0.16		
shape:jitter:shimmer	3.9	0.05*				
shape:jit:pitch:shim					6.4	0.01*

indicate that more subjects are needed to come up with more confident conclusions.

When we compare two emotions then we can compile a table with predictors as in table 6.8, where we left out the F values. Remarkably, hinting to tables 6.9 and 6.10, the two statistical packages come up with practically the same values, and exactly the same significances. The table confirms that pitch is very decisive in discriminating emotion from no emotion. Shape does not play a role anymore when discriminating fear from hot anger; it does for the more aroused emotions (if we may make this association) compared to the least aroused emotions. Jitter and shimmer demonstrate to be significantly discriminative for hot anger against the less aroused

Table 6.10: *Dependent variable degree is fitted for the independent variables shape, pitch, jitter and shimmer. Horizontal are the emotions, which are kept constant for each analysis. Statistics analyzed using Statistica. (* denotes significance)*

	p for degree per emotion					
	Cold Anger		Fear		Hot anger	
Predictor	F	p	F	p	F	p
shape	1.1	0.35	1.2	0.31	3.4	0.018
pitch	2.5	0.04*	23.7	< 0.000001*	2.7	0.03*
jitter	2.1	0.12	2.4	0.09	9.3	0.0001*
shimmer	1.6	0.20	0.5	0.61	0.5	0.64

Table 6.11: *Confusion matrix for subjects choice of emotion. BS ... BtB are abbreviations for the names of the eight participating subjects.*

	Observations								
Emotion	BS	SB	RW	JB	RS	DT	CtB	BtB	Totals
Hot Anger	1	0	76	43	56	18	1	194	389
Fear	4	16	95	72	292	105	42	140	766
Cold Anger	5	0	121	182	161	62	4	178	713
Neutral	530	524	248	243	31	355	493	28	2452
Totals	540	540	540	540	540	540	540	540	

emotions.

To gain a better understanding on what the effect of each predictor is on the perceived emotion, we determine what happens with the dependent variable degree, when we consider a single emotion. If for some emotion a strong correlation with degree is found, we can conclude which predictors are decisive for which emotion. Results are shown in tables 6.9 and 6.10. The first table uses R and the second uses Statistica to calculate the statitistics. Under Statistica categorical emotion labels were used, but under R emotions are rendered into numericals. It was expected that there would be no difference in this case, since only one emotion value was used. Using Statistica it was not possible to calculate predictor interactions. Strikingly, shape does not play a role in the second analysis, but it clearly does in the first. There is a trend observable, however, showing that shape does become more of a predictor for the more aroused emotions. Pitch plays a role in all three emotions (except for cold anger using R), but apparently fear is mostly associated with high pitch. Jitter becomes significant for fear and hot anger. Evidently, shimmer alone is never a significant predictor, but interactions with it are.

It is of course of importance to know how our subjects differed in their choice of emotions. Since only eight subjects are used in this experiment, we can plot a confusion matrix showing the choices of each subject. As table 6.11 shows, subjects did not agree unanimously. Skipping subjects BS and SB (symmetry is a coincidence) does not change much for the statistics as far as tables 6.4 – 6.6 are concerned. Only for degree significances becomes stronger. All emotions taken apart, like in table 6.10, while leaving BS and SB out of the data does not change much either: for cold anger pitch is significant with an improved $p = 0.02$, for fear jitter improves to $p = 0.05$, and for hot anger a slight improvement for shape, $p = 0.017$, is obtained.

Table 6.12: *Comparing pairs of levels of degree, leaving other levels of degree out of the data.*

	p for pairs of degree					
Predictors	1–2	1–3	1–4	2–3	2–4	3–4
shape			0.0001		0.0007	0.002
pitch	0.01	0	0	0	0.0007	
jitter			(0.06)	(0.08)		(0.1)
shimmer	0.04	0.03				
shape:pitch				0.03		(0.05)
pitch:shimmer				(0.06)		
observations	903	858	567	1301	1010	965

Degree discrimination

For degree we also put together a table comparing the influence of predictors for pairs of degree (table 6.12). Shape apparently is discriminative for high value of degree, at least for the shapes we selected. For the highest two values of degree pitch ceases to be discriminative, however, our ears being nonlinear devices, we have to note we did not account for relative distances of pitch frequencies. Also our data set may be too small. So, we expect that pitch is discriminative for degree for a broad range of frequencies. Shimmer plays a role when pitch is still low, early evidence of pitch being a stronger predictor for the shimmer values we tested.

A final observation: pitch, jitter and shape determined the realism of the sounding of our synthesized vowels.

Effect of F_0 approaching F_1

People can use the fact that when they increase pitch frequency, such that F_0 is close to the first formant frequency, F_1, loudness also increases. Nett result is that they gain a higher effect with less effort. Effect on transmitting a message is better speech intelligibility in a noisy environment or more focus on the fact that the messenger is in a aroused state of mind. Schutte (1999) studied this kind of effects[5] with singers. In our data there are samples for which pitch frequency is close to F_1 for the vowels /e/ and /u/. They have a F_1 of respectively 442 Hz and 385 Hz (table 6.1). This fact asked for a closer look at which vowels received the most emotion labels by our subjects. One way to do this is to count the number of different emotions for each vowel. However, we did not find any evidence for an increase of emotion or degree related to pitch frequency approaching the first formant frequency.

[5] Actually, his study included the effect of F_1 and F_2 coming close together. Since we used fixed values for the formants in our experiment, we were inspired to look at F_0 approaching F_1.

Learning effect

To investigate learning effects one could split up data in groups and apply statistics for each of the groups and compare them. Here we first inspected bivariate histograms showing the number of observations for each emotion compared to item number. Every sound fragment each subject listened to was randomly selected and we therefore assume that no bias was present in the presentation of items. If there is no learning effect then on average the number of observations should be constant. The bivariate histogram hinted towards a learning effect for fear. Subjects seemed to rate early fragments more frequently as fear than later fragments. Mean plots showed more evidence for this effect. We compared all six possible combinations of emotion labels and inspected the evolving of the mean emotion during the experiment grouped for all subjects. For the combinations with fear the mean always tended to decrease, while for the other combinations this was not true, except for the combination cold anger and hot anger; here also a (small) decrease in mean was

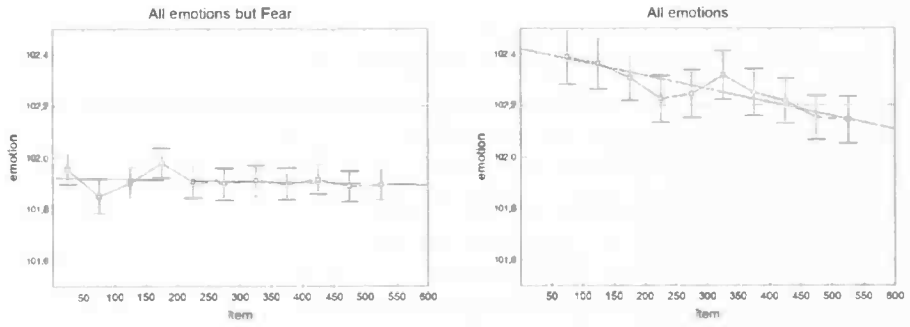


Figure 6.6: *Mean plots of emotions with 95% confidence intervals, showing evidence of a learning effect. (a) Mean of all emotions but fear. (b) Mean of all emotions, fear included. Horizontal item numbers are given (1 ... 540)*

observed. Figure 6.6 depicts the learning effect showing mean emotion (in arbitrary units) without and with fear.

6.3 Discussion

We found that it is not trivial to synthesize natural vowels. Our experiment shows that some vowels are more easy to define, so to speak, than others. The vowel /a/ perceptually gave good results using six formant frequencies and bandwidths only. For other vowels, here the /e/ and /u/, subjects did not agree that convincingly with our intentions. But, since we, for instance, virtually left out all prosodic information, we cannot be sure about to what extend this vowel confusion is related to vowel identity. (As noted before, confusion can be explained by the vowel triangle and increasing F_0.) Subjects did agree with our assumption of the sound fragments being a vowel however. So, basically, we succeeded in synthesizing vowels.

Based on our analysis of the results our four emotions can be ordered, in increasing content of arousal, as neutral < cold anger < fear < hot anger. This result is predicted by FEELTRACE and forms evidence for the capability of the vocoder to add emotional content to synthetic vowels. Our hypothesis (page 53) seems to be correct in the context of our experiment.

Of course, we would have been more satisfied if results showed higher rates for realism of the data. But, shape proved to be a very important predictor for realism, as was hypothesized, indicating the importance of focus on the perception side of the speech chain. Moreover, realism increases –very profoundly– when prosody is introduced, reason for the introduction of our pitch shape function. Eventually, future experiments should include effects of prosody.

In 43% of the fragments emotion was perceived. Fear and hot anger show a clear positive trend with shape. But this does not help us in classifying emotions rather than detecting them. That is, not if we group emotions into categories, as we obviously do. To come to a classification we have to relate the perceived degree (of an emotion) to the predictors for each emotion individually. We can conclude that when emotions with a high arousal become more intense, spectral moment does increase. Also, our first experiment indicates that this is more apparent for the more 'extreme' emotions, e.g. shape is more strongly positive related to increasing hot anger than to increasing fear. On the other hand, pitch is very strongly positively related to degree of fear, more than it is for hot anger. The latter seems to be better predicted by the amount of jitter. There is, however, overlap for the predictors, especially for fear and hot anger. So, there seem to be possibilities to come to classification, in future, when more predictors are compared and used to compute probabilities indicating what emotion and degree to expect.

To avoid some biasing (and perhaps learning) effect, we wonder if a future experimental setup should employ the method of comparing two sound fragments. This method is used to find just noticeable differences (JND) and could be adapted to find the impact of the various predictors. An advantage is that a subject has to focus on the difference of fragments, rather than being preoccupied with finding regularities in the presented data and mapping it to the designer options.

One final issue about methodological problems we like to mention. It is questioned by Scherer (1986) whether decoding accuracy results might be spurious? It is possible that methodological artifacts are responsible for the high accuracy percentage obtained in decoding studies (ibid). Subjects might be able to *guess* or use *rules* to exclude possibilities to direct their performance. Some evidence of the evolving of rules was found after interviewing the subjects. This behavior may be due to the fact that subjects are presented only a small number of response alternatives. On the other hand, subjects were presented a large set of fragments, which took up to about one hour of time to complete, and subjects afterwards responded that they thought these fragments were genuine, however manipulated, vowel fragments. It might be that the duration of the experiment compensated for some of the effects mentioned.

Chapter 7

Conclusions and future work

To climb steep hills requires slow pace at first.

William Shakespeare 1564–1616

Although classification of aggression necessitates workable definitions of levels and categories of emotion, and in this work we only loosely defined a set of emotions, it is our opinion that further research on the production side of speech will come up with more robust definitions for cues. Classification is difficult and maybe impossible using a sound signal alone. So, we probably should first determine the amount of knowledge, kinds of heuristics and redundancy (other modalities) humans use to do their magnificent job of classification. More robust emotion detection might be possible though, using the sound signal, without too much context information. The fact that people demonstrate loss of control over their speech production process (e.g. vocal folds) opens the door to the derivation of cues related to this loss. In this work a simple definition of jitter, for instance, yielded promising results. Jitter is, in our humble opinion, worth further exploring in the context of aggression classification.

It comes to mind that a next experiment should use a higher resolution for shape as a predictor, and probably try to extract some descriptive terms from it. But even though we probably used a rather course set of cue definitions, the outcome of the experiment confirms our hypothesis, i.e. increasing spectral content, or spectral moment, is perceptually related to a more aroused behavior. So are jitter and shimmer and pitch. But to be able to firmly draw some lines, as to where some cue becomes more indicative for some emotion, more experiments have to be set up. Especially since pitch came out to be very strongly correlated to the 'aroused emotions', we should more carefully chose parameter values in a next experiment. Moreover, since we used a rather simple definition for jitter (and shimmer) we should try to come up with a more realistic definition of it. It is evident, however, that the vocoder is of use in generating a synthetic database of speech fragments such that these fragments can be used to test the relevance of emotion cues (e.g. in aggression).

As discussed before, our experiment indicates to the successful use of speech production related cues to come to a better detection or classification of emotions and degree of emotions even. This is based on the observation that the predictors, as used in the experiment, did not constitute equivalent models for the emotions. Using a set of cues to collect evidence seems to be a viable approach to build a working emotion classifier. It may use Bayesian techniques or Hidden Markov Models.

Prosody was not considered in our experiment. A lack of prosody primarily implies a lack of naturalness. But, paralinguistic aspects contain important information and thus are important for humans in decoding a message.

On the longer term, higher level processing should be considered. We find systems as used in cognitive science, like ACT-R[1], very inspiring. It might be advantageous to develop an architecture for simulating and understanding human speech production and perception. This way the collecting of evidence and the concept of information feedback can be implemented and theorized. Researchers could profit by the fact that they could work on a aspect of the whole and could evaluate it separately and, subsequently, in relation to other modules comprising the architecture or theory, thereby focusing on, e.g., aggression classification.

The main objective of this thesis to develop, implement and evaluate a vocoder is, in our opinion, successfully achieved. The vocoder promises to be of use in research and the model can easily be implemented and provided with a graphical user interface (GUI), of which apendix C shows an example. Having regard to the research question, we focused on spectral features directly related to the speech production process. Of course this is directly related to the perception of emotions, but it is our belief that this relation is not a one-to-one relation. What is measured on the receiver side of the speech chain, may relate back to different sources on the production side. That is, as far as current cues are concerned. So, further unraveling of spectral content and relating this to the sound source(s) is needed to come to correct inferring and conclusions on the cochlea side.

A question is how far nonlinear methods will bring us? The human hearing system does behave linearly for most part, but nonlinearities play a very important role on the speech production side. Very interesting methods are being invented to investigate these nonlinearities. A lot of literature is advocating nonlinear research. In future work we would like to employ the Teager Energy Operator to do pattern recognition on speech modulations. It is true that there is a great deal of variability in the human voice when we look at jitter and shimmer, but these natural variation might be different in nature from jitter and shimmer induced by the changing arousal of a speaker. As noted before in this text, we expect that there is a need to come up with more, or even better, criteria to differentiate between spectral qualities or quantities.

Also the proceedings in chaos theory should not be forgotten. Period doubling is present and detectable in sound recordings, for instance, and chaos theory might be the tool to describe these phenomena. We are aware of the fact that these methods may not have any direct practical use, e.g. for Sound Intelligence, but we should keep an open mind!

In this work we used a formant synthesizer. The amount of parameters used is quit large. In future work we could try to reduce the number of parameters by combining and mapping to higher level descriptions. An interesting topic is to use physical models to produce speech. They will require to solve Navier-Stokes equa-

[1]Find the ACT-R homepageat http://act-r.psy.cmu.edu/.

tions, and the like, which is a computationally expensive task. But, predictive power of such models is expected to be much more great than that of the linear formant synthesis model.

Appendix A

Interaction response tables

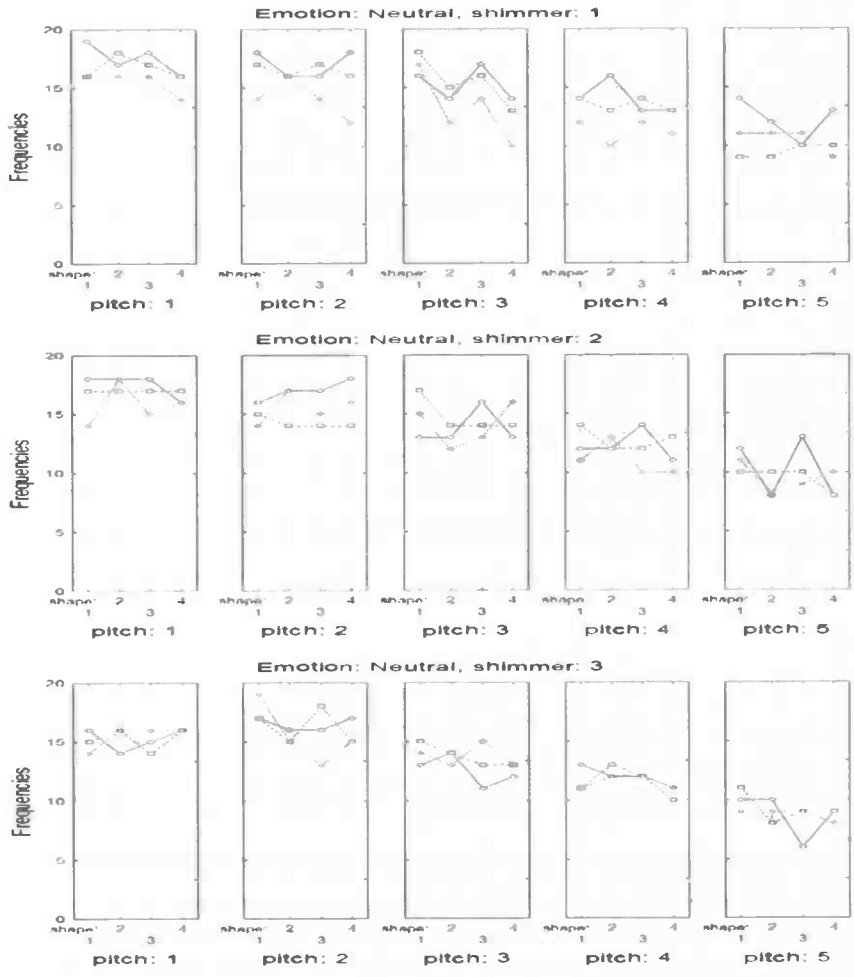


Figure A.1: *Interaction plots: effect of shape, pitch, jitter and shimmer on emotion* `Neutral`. *Every plot shows five sub plots, one for each pitch value. On the horizontal axes of every sub plot shape increases, and every line in a sub plot indicates a different value of jitter (blue solid lines with round markers for jitter = 1, red dashed lines and boxed markers for jitter = 2, and green dotted lines with triangular markers for jitter = 3). From top to bottom plot shimmer is respectively 1, 2, and 3.*

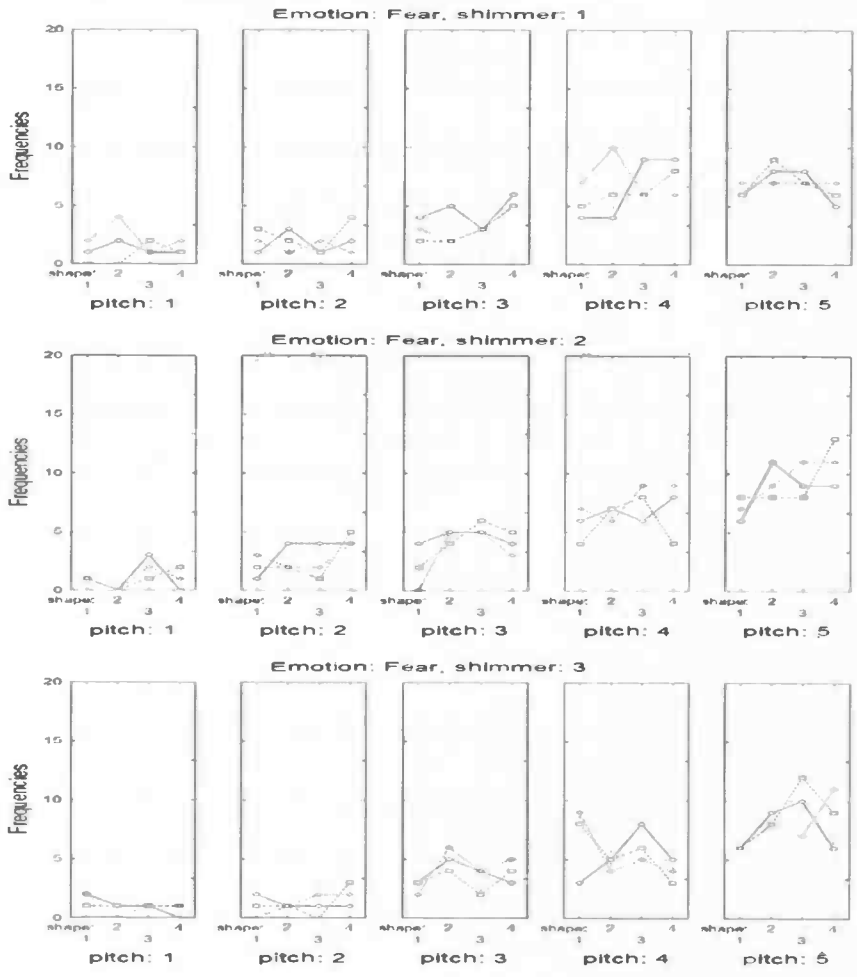


Figure A.2: *Interaction plots: effect of shape, pitch, jitter and shimmer on emotion* `Fear`. *Every plot shows five sub plots, one for each pitch value. On the horizontal axes of every sub plot shape increases, and every line in a sub plot indicates a different value of jitter (blue solid lines with round markers for jitter = 1, red dashed lines and boxed markers for jitter = 2, and green dotted lines with triangular markers for jitter = 3). From top to bottom plot shimmer is respectively 1, 2, and 3.*

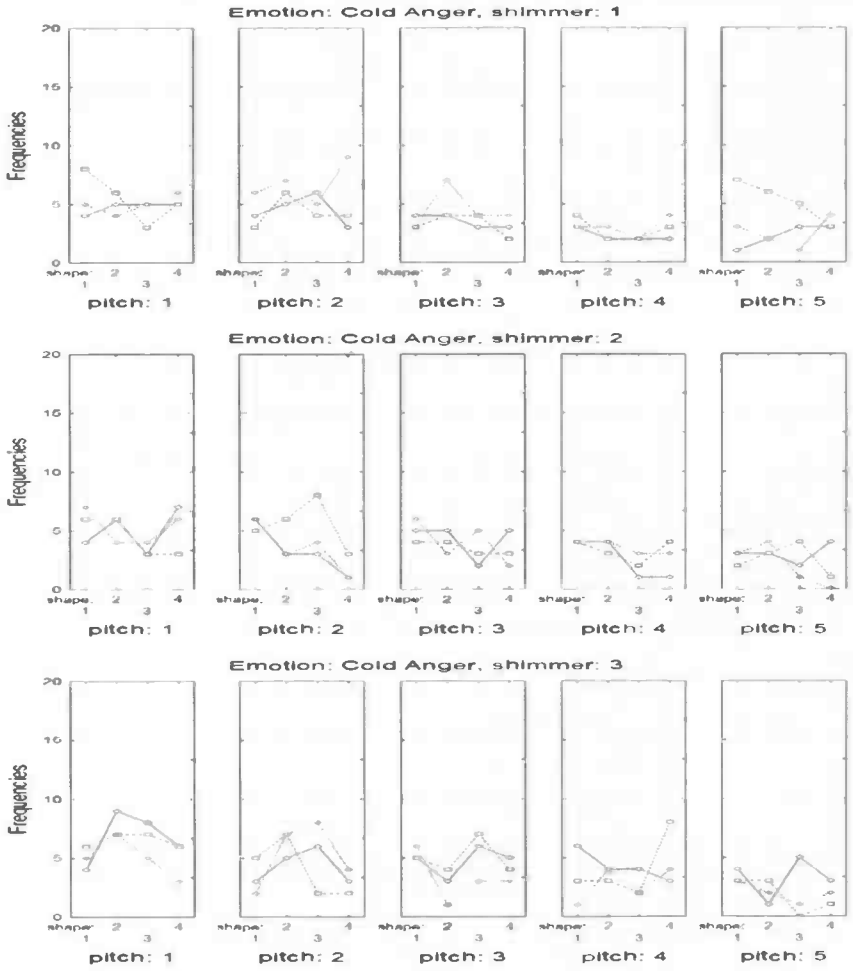


Figure A.3: *Interaction plots: effect of shape, pitch, jitter and shimmer on emotion* `Cold Anger`. *Every plot shows five sub plots, one for each pitch value. On the horizontal axes of every sub plot shape increases, and every line in a sub plot indicates a different value of jitter (blue solid lines with round markers for jitter = 1, red dashed lines and boxed markers for jitter = 2, and green dotted lines with triangular markers for jitter = 3). From top to bottom plot shimmer is respectively 1, 2, and 3.*

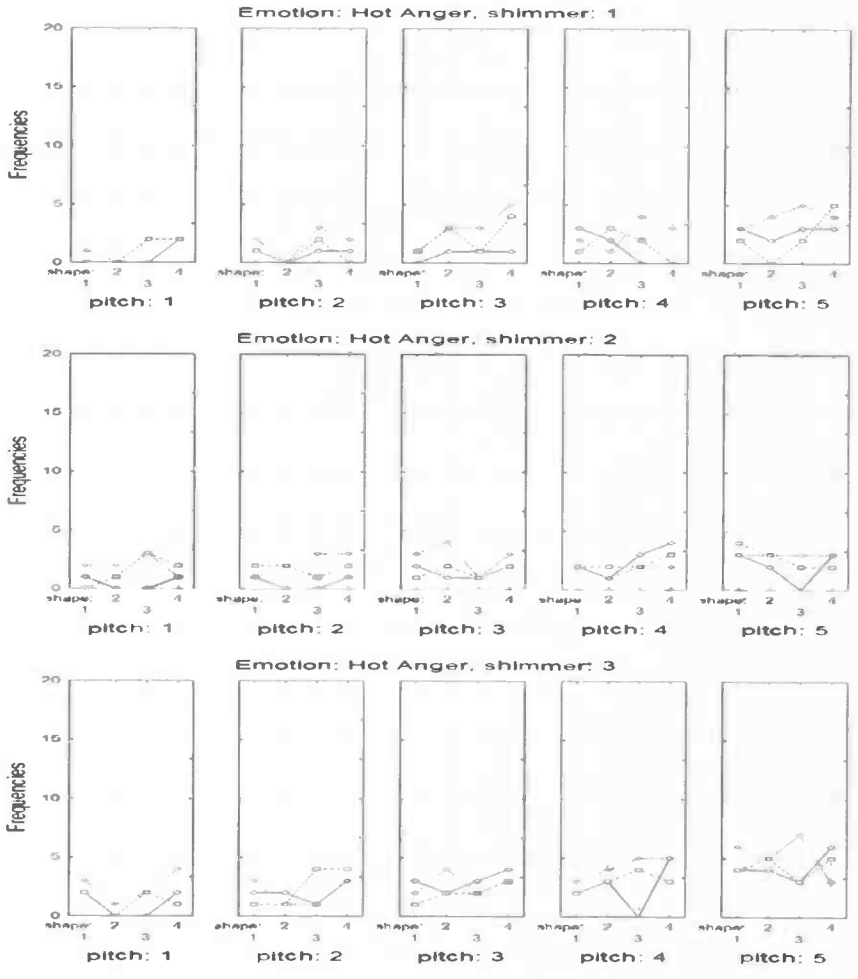


Figure A.4: *Interaction plots: effect of shape, pitch, jitter and shimmer on emotion* `Hot Anger`. *Every plot shows five sub plots, one for each pitch value. On the horizontal axes of every sub plot shape increases, and every line in a sub plot indicates a different value of jitter (blue solid lines with round markers for jitter = 1, red dashed lines and boxed markers for jitter = 2, and green dotted lines with triangular markers for jitter = 3). From top to bottom plot shimmer is respectively 1, 2, and 3.*

Appendix B

Translations to Dutch of some terminology

English	Dutch
alveola	tandkas(sen)
arytenoid cartilages	bekerkraakbeentjes
cartilage	kraakbeen
cricoid (or: cricoid cartilage)	ringkraakbeen
diaphragm	middenrif
epiglottis	strotklepje
glottis	stemspleet
hyoid bone	tongbeen
larynx	strottenhoofd
mandibula	onderkaak
palate, palatum	verhemelte
pharynx	keelholte
thyroid	schildkraakbeen
trachea	luchtpijp
velum	zacht verhemelte

Appendix C

Vocoder GUI

Below an example of a graphical user interface (GUI) for the vocoder is shown. A speech segment can be divided into a number of frames of t_{frame} ms. Frequencies for fundamental frequency, F_0, formants, e.g. $F_1 \ldots F_5$, and bandwidths can be set for every frame. Also relative source amplitude can be set per frame. Several drawing techniques can be implemented to assist the user. For instance, using splines it is possible to define a smaller set of points and then draw a fitting line through them. The inset graph gives the user an idea of the effect of the current formant frequency settings. Here it does not show what would be measured when simulating the vocal tract. Parameters can be imported from and exported to Matlab. Also Matlab functionality can be made available via code libraries.

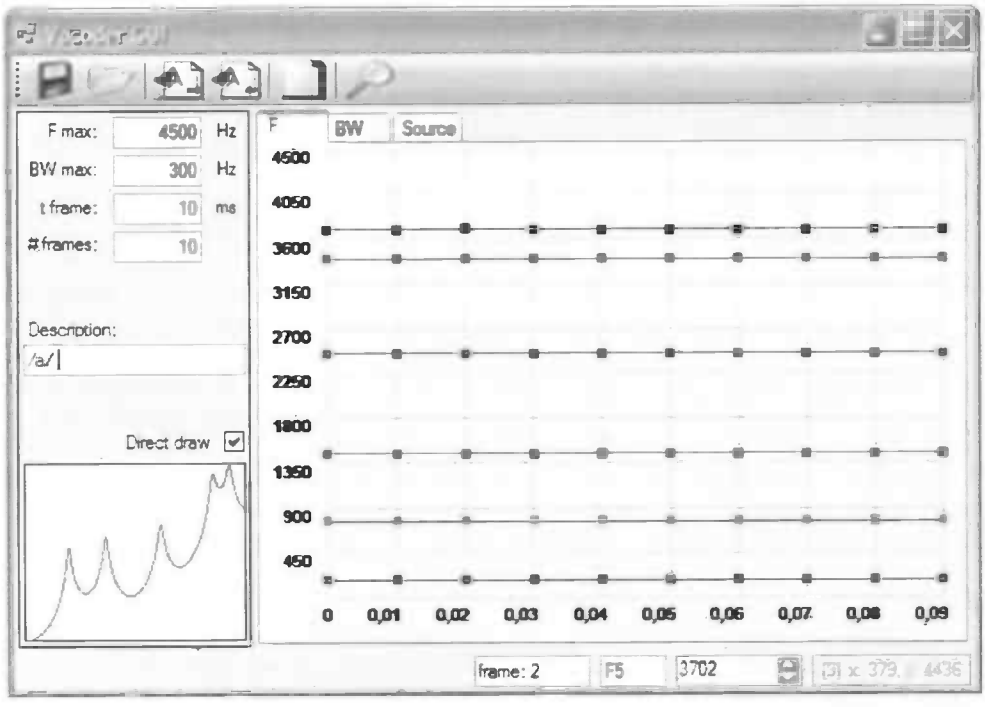


Figure C.1: *An example of a vocoder GUI.*

Bibliography

[1] Adank, P., van Hout, R., and Smits, R. (2004). An acoustic description of the vowels of northern and southern standard dutch. *Journal of the Acoustical Society of America*, 116(3):1729–1738.

[2] Alonso, J. B., de Maria, F. D., Travieso, C. M., and Ferrer, M. A. (2005). Using nonlinear features for voice disorder detection. *NOLISP-2005*, pages 94–106.

[3] Andringa, T. C. (2002). *Continuity perserving signal processing*. PhD thesis, Rijks*universiteit* Groningen.

[4] Asogawa, S. and Akamatsu, N. (1999). A new model for the source of vowels based on the vortex sound. *Inf. Sci.*, 116(2-4):165–176.

[5] Bakshi, R. (2004). Klatt's speech synthesizer: A case study for hardware/software codesign. Master's thesis, Indian Institute of Technology Delhi.

[6] Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

[7] Baron, J. and Li, Y. (2004). *Notes on the use of R for psychology experiments and questionnaires*. Department of Psychology, University of Pennsylvania.

[8] Boersma, P. (1995). Interaction between glottal and vocal-tract aerodynamics in a comprehensive model of the speech apparatus. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 2, pages 430–433, Stockholm.

[9] Childers, D. G. and Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5):2394–2410.

[10] Chowning, J. (1999). Perceptual fusion and auditory perspective. In Cook, P. R., editor, *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics*. MIT Press, Cambridge, Massachusetts, USA.

[11] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000). "Feeltrace": An instrument for recording perceived emotion

in real time. In *Proceedings of the ISCA Workshop (ITRW) on Speech and Emotion: Developing a Conceptual Framework*, pages 19–24, Queen's University of Belfast, N.I.

[12] De Vries, M. P., Hamburg, M. C., Schutte, H. K., and Verkerke, G. J. (2003). Numerical simulation of self-sustained oscillation of a voice-producing element based on navier-stokes equations and the finite element method. *The Journal of the Acoustical Society of America*, 113(4):2077–2083.

[13] De Vries, M. P., Schutte, H. K., Veldman, A. E. P., and Verkerke, G. J. (2002). Glottal flow through a two-mass model: Comparison of navier-stokes solutions with simplified models. *The Journal of the Acoustical Society of America*, 111(4):1847–1853.

[14] Dimitriadis, D. and Maragos, P. (2001). An improved energy demodulation algorithm using splines. In *International Conference on Acoustics, Speech, and Signal Processing, 2001.*, volume 6, pages 3481–3484, Salt Lake City, UT, USA.

[15] Doval, B. and d'Alessandro, C. (1997). Spectral correlates of glottal waveform models: an analytic study. In *Proc. ICASSP '97*, pages 1295–1298, Munich, Germany.

[16] Fant, G. (1970). *Acoustic Theory of Speech Production*. Mouton, The Hague - Paris, 2^{nd} edition.

[17] Flanagan, J. L. (1957). Note on the design of "terminal-analog" speech synthesizers. *Journal of the Acoustical Society of America*, 29(2):306–310.

[18] Hagmüler, M., Rank, E., and Kubin, G. (2004). Can stress be observed by analyzing the human voice? In *3rd Eurocontrol Innovative Research Workshop*, Graz University of Technology, Austria.

[19] Hogg, R. V. and Thanis, E. A. (2001). *Probability and Statistical Inference*. Prentice Hall, 6^{th} edition.

[20] Huisman, M. (2004). Akoestische effecten van emoties in spraak: De waarneming van verbale agressie. Master's thesis, Rijks*universiteit* Groningen.

[21] Jesus, L. M. T. d., Vaz, F., and Principe, J. C. (1997). An implementation of the Klatt speech synthesiser. *Revista do Detua*, 2(1).

[22] Junqua, J.-C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 1:510–524.

[23] Keller, E. (2005). The analysis of voice quality in speech processing. In Chollet, G., Esposito, A., Faundez-Zanuy, M., and Marinaro, M., editors, *Nonlinear Speech Modeling and Applications: Advanced Lectures and Revised Selected Papers*, volume 3445, pages 54–73. Springer Berlin/Heidelberg.

[24] Klasmeyer, G. (2000). An automatic description tool for time-contours and long-term average voice features in large emotional speech databases. *Speech Emotion 2000*, pages 66–71.

[25] Klatt, D. and Klatt, L. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

[26] Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995.

[27] Kvedalen, E. (2003). Signal processing using the Teager Energy Operator and other nonlinear operators. Master's thesis, University of Oslo, Norway.

[28] Lemmetty, S. (1999). Review of speech synthesis technology. Master's thesis, Helsinki University of Technology.

[29] Little, M., McSharry, P., Moroz, I., and Roberts, S. (2006). Testing the assumptions of linear prediction analysis in normal vowels. *The Journal of the Acoustical Society of America*, 119(1):549–558.

[30] Liu, X., Povinelli, R. J., and Johnson, M. T. (2003). Vowel classification by global dynamic modeling. In *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, pages 111–114, Le Croisic, France.

[31] Mann, I. and McLaughlin, S. (1998). A nonlinear algorithm for epoch marking in speech signals using poincar maps. In *Procedings of the 9th European Signal Processing Conference EUSIPCO*, volume 2, pages 701–704.

[32] Maragos, P., Dimakis, A. G., and Kokkinos, I. (2002). Some advances in nonlinear speech modeling using modulations, fractals, and chaos. In *Proceedings of International Conference on Digital Signal Processing*, volume 1, pages 325–332, Santorini, Greece.

[33] Maragos, P., Kaiser, J., and Quatieri, T. F. (1993). Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, 41(10):3024–3051.

[34] Menzer, F. (2004). Modeling transient behaviour in vocal fold vibration using bifurcating nonlinear ordinary differential equation systems. Master's thesis, Swiss Federal Institute of Technology, Lausanne.

[35] Mertens, P. (2002). Synthesizing elaborate intonation contours in text-to-speech for french. In Bel, B. and Marlien, I., editors, *Proceedings of the Speech Prosody 2002 conference 11-13 April 2002*, pages 499–502, Aix-en-Provence: Laboratoire Parole et Langage.

[36] Movellan, J. R. (2002). Tutorial on gabor filters. `http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf`.

[37] Mozziconacci, S. J. (1998). *Speech Variability and Emotion: Production and Perception*. PhD thesis, Technical University Eindhoven.

[38] Murray, I. R., Baber, C., and South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication*, 20(1-2):3–12.

[39] Niessen, M. (2004). Speaker specific features in vowels. Master's thesis, Rijks*universiteit* Groningen.

[40] Ohala, J. J. (1978). Production of tone. In Fromkin, V. A., editor, *Tone: a linguistic survey*, pages 5–39. Academic Press, New York.

[41] Plutchik, R. (1980). *Emotion: a psychoevolutionary synthesis*. Harper & Row, New York.

[42] Pols, L. C. W., Tromp, H. R. C., and Plomp, R. (1973). Frequency analysis of dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53(4):1093–1101.

[43] Rietveld, A. and Van Heuven, V. (2001). *Algemene fonetiek*, chapter 1; 10, Wat is Fonetiek? Uitgeverij Coutinho b.v.

[44] Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2B):583–590.

[45] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165.

[46] Schoentgen, J. (1990). Non-linear signal representation and its application to the modelling of the glottal waveform. *Speech Communication*, 9:189–201.

[47] Schoentgen, J. (2003). Shaping function models of the phonatory excitation signal. *Journal of the Acoustical Society of America*, 114(5):2906–2912.

[48] Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, Institute of Phonetics, Saarland University.

[49] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In Dalsgaard, P., Lindberg, B., and Benner, H., editors, *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, volume 1, pages 87–90, Aalborg. Kommunik Grafiske Losninger A/S.

[50] Schutte, H. K. (1999). Fysiologie van de stemgeving. In Peters, H. F. M., Bastiaanse, R., van Borstel, J., Dejonckere, P. H. O., Jansonius-Schultheiss, K., van der Meulen, S., and Mondelaers, B. J. E., editors, *Handboek Stem-, Spraak-, Taalpathologie*, volume 10, chapter A3.1.1, pages 1–37. Bohn, Stafleu, Van Loghum, Houten.

[51] Shadle, C. H., Barney, A., and Davies, P. (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. ii. implications for speech production studies. *The Journal of the Acoustical Society of America*, 105(1):456–466.

[52] Stevens, K. N. (2002). Toward formant synthesis with articulatory controls. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pages 67–72.

[53] Stewart, I. (1997). *Does God Play Dice?* Penguin Books Ltd, 2nd edition.

[54] Titze, I. (1994). *Principles of Voice Production*. Prentice Hall, Englewood Cliffs, NJ.

[55] Toivanen, J., Seppänen, T., and Väyrynen, E. (2003). Creation and utilization of the MediaTeam Emotional Speech Corpus. In *Proc. Corpus Linguistics 2003*, volume 16 of 2, pages 791–799, Lancaster, UK.

[56] Van Dinther, R. (2003). *Perceptual aspects of voice-source parameters*. PhD thesis, Technische Universiteit Eindhoven.

[57] Van Dinther, R., Veldhuis, R., and Kohlrausch, A. (2005). Perceptual aspects of glottal-pulse parameter variations. *Speech Communication*, 46:95–112.

[58] Veldhuis, R. (1998a). A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103(1):566–571.

[59] Veldhuis, R. N. J. (1998b). The spectral relevance of glottal-pulse parameters. In *Proc. ICASSP*, volume II, pages 873–876.

[60] Verhoeven, J. and van Bael, C. (2002). Akoestische kenmerken van de nederlandse klinkers in drie vlaamse regio's. *Taal en tongval*, 54(1):1–23.

[61] Zhou, G., Hansen, J. H. L., and Kaiser, J. F. (2001). Ninlinear feature based classification of speech under stress. *IEE Transactions on Speech and Audio Processing*, 9(2):201–216.