rijksuniversiteit
groningen

Development of a tool for Soundscape Annotation

What do we hear when we listen?

R. van der Linden
March 2011

Master Thesis
Submitted for the degree of Master of Science in Artificial Intelligence

Artificial Intelligence - Auditory Cognition Group
Dept of Artificial Intelligence,
University of Groningen, The Netherlands

Primary Supervisor: Dr. T. Andringa, University of Groningen
Secondary Supervisor: Dr. D.J. Krijnders, University of Groningen, INCAS^3

Contents

Abstract

In the developing field of automatic sound recognition there exists a need for well-annotated training data. These data currently can only be gathered through manual annotation; a time-consuming and sometimes tedious task. How can a software tool support this task? The objective of this master's project is to develop and validate a tool for soundscape annotation. Furthermore we assess the strategies that subjects employ when annotationing a real-world sound recording. In an experiment with untrained participants, annotations were collected together with user data (keystrokes and mouse clicks) that provide insight in the strategies subjects employ to achieve the annotation task. Dividing attentional resources over the time span of the recording is an important aspect of the task.

Soundscape annotation, the process of annotating a real-world sound recording, can be seen as a special case of 'everyday listening' (Gaver). When annotating an audio recording offline (as opposed to reporting auditory events 'in vivo') the subject lacks context knowledge, but offline annotation also opens new possibilities for the listener, for example to listen to the same sound event more than once. These differences have implications for the task and ultimately bring the question to mind: what makes a 'good' annotation?

Chapter 1

Introduction

Sound is everywhere around us. Apart from deserted areas and well-isolated rooms, everywhere humans go they perceive sound. From a physical perspective, the sound waves entering the ear form a seemingly unstructured mess of vibrating air at different frequencies; humans however have the capacity to structure mess into meaningful elements, analyze those elements and may even be said to understand the world through sound. Composer and environmentalist Shafer proposed the term *soundscape* (Schafer 1977) to describe the sonic environment as a human perceives; the subjective experience resulting from the auditory input, so to say. A soundscape must be seen as the auditive counterpart of what perceiving a *landscape* is in vision; a soundscape basically is the perceived sonic environment.

This masters thesis is centered around the idea of *annotating* soundscapes. An *annotation* to (part of) a source of information is a summary: as a description, often in text, the annotation briefly describes the contents of the source, relevant to some task or goal. A soundscape can be perceived in its natural environment; in vision expanded in this thesis, the characteristic information contained in a soundscape can also be captured in a (digital) recording.

But why would one want to annotate soundscapes? One reason is that the application and storage of digital recordings is ever increasing, and therewith the demand for a method to enrich sound recordings with detailed descriptions and content information increases. Annotations describing sound sources or events provide those descriptions.
Another application is the practice of testing and training automatic sound recognition algorithms: this demands a precise and accurate annotated database of sound recordings. For this application often a large body of training data is needed, but well-annotated databases are not yet available for this purpose because acquiring annotations is expensive and time-consuming.

The project described in this thesis seeks to develop a method for (human) soundscape annotation using a software tool to improve annotation speed without paying a toll on accuracy and descriptiveness. If successful, such a method will help to make more well-annotated soundscapes available that may enable the field of sound recognition to increase the performance of automatic recognizers significantly. More on the potential applications can be found in section 1.0.1 below.

There is another reason why soundscape annotations made by humans are worthwhile: **Studying the process of human soundscape annotation may reveal fundamental aspects of human audition.** The setting provides a platform to research auditory perception in the special case of 'listening to annotate': listening closely and reporting what you hear in a soundscape. When looking at the soundscape annotation task from this cognitive point of view, scientific questions arise on the nature of hearing and listening in this more or less artificial setting. One may ask: How does the task influence the resulting annotations? How does listening to a reproduced soundscape relate to perceiving a soundscape the 'real world'? How is the absence of a large part of the context (namely the other perceptual experiences) reflected in the resulting annotations? By researching auditory perception in this domain knowledge may be gained about auditory perception in general. Therefore this thesis extensively reviews literature on auditory perception and attempts to link findings on general auditory perception to the special case of listening in an annotation task.

The layout of this thesis is as follows: In the background chapter we review a range of scientific disciplines that are connected to the topic of (semi-automatic) soundscape annotation. Next, the development of a dedicated software tool for (human) soundscape annotation is described, together with the design requirements and choices. This tool was tested in an experiment that is described in chapter 4. In this experiment 21 subjects performed a semantic annotation task under different time constraints. Usability information was collected while the subjects annotated a sound recording, and afterwards a survey was held among the participants. Chapter 5 presents the results of this experiment. In the discussion in chapter 6 these results are interpreted; together with the resulting annotations the data was analyzed to see how the tool performs, what strategies the subjects exhibit in carrying out their annotation task and how the resulting annotation sets differ between subjects and conditions.

In the remainder of this introductory chapter the concept and applications of annotating soundscapes will be introduced.

The previous section already pointed out that the demand for well-annotated soundscapes is not isolated, but is closely linked with the digitalization of information throughout society. Ever more information is stored digitally in our modern world; one can think of broadcasts such as radio and television programs that are sent out in a digital encoding. Another example is security and surveillance appliances digital in which recognition and storage may take place, see (Van Hengel and Andringa 2007) for a successful implementation of such a system. Once an audio recording is made and stored on a recording device or harddisk, there will often exist a need for describing the contents of the recording, depending on the purpose of the recording. One option is to *tag* the recording with short semantic descriptions that describe the contents: in the above example of an urban soundscape, the tags attached to the recording could be:

$$\{traffic, speech, constructionwork\}$$

These semantic descriptions indicate that the sound events of these three categories are contained in the recording. However, when the recording spans minutes or hours it is more

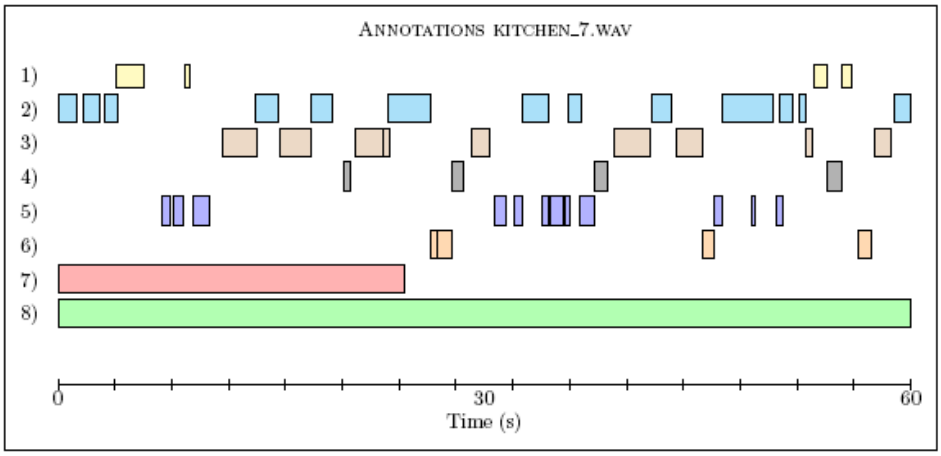

Figure 1.1: *Graphical representation of the annotation of a recording of doing the dishes. Every line represents a different class: 1) Splashing, 2) Scrubbing dishes, 3) Rummaging through the sink, 4) Dropping cutlery, 5) Clanging cutlery, 6) Water drips, 7) Fridge, 8) Boiler pump. From (Grootel et al. 2009)*

useful to also store order and timing information. We might want to describe more precisely *where* in the sound file these sound sources occur, i.e. store the exact 'location' in time of the source within the recording. Depending on the application, frequency information might also be included in the annotation. To illustrate how this information could be represented graphically on a timeline, an example of the annotation of a soundscape recording from a recreational area, containing a mixture of sources, is given in figure 1.1.

The soundscape recordings this project seeks to annotate are recorded in *real world, uncontrolled environments* where no artificial constraints were placed on the sounds events that were captured. A more detailed discussion of *real world sounds* is given in section 2.2. In creating these recordings, no actions were taken that could influence the recordings. Further information on the recording method can be found in section 4.1.

Soundscape annotations can take on different forms, each form having its own advantages and disadvantages. A number of options is discussed in the next chapter. In the view expanded in this thesis, the (human) annotation task consist of the following actions:

1. Listening to the sound recording to hear and recognize sound sources,
2. Indicate the point or temporal interval for which that sound sources was perceived,
3. Attaching a semantic description to the annotated part of the recording.

1.0.1 Applications for annotated sound recordings

Collecting annotations for sound files through human input is costly. Are these annotations worth all the effort? Where might annotated soundscapes find their application? The main

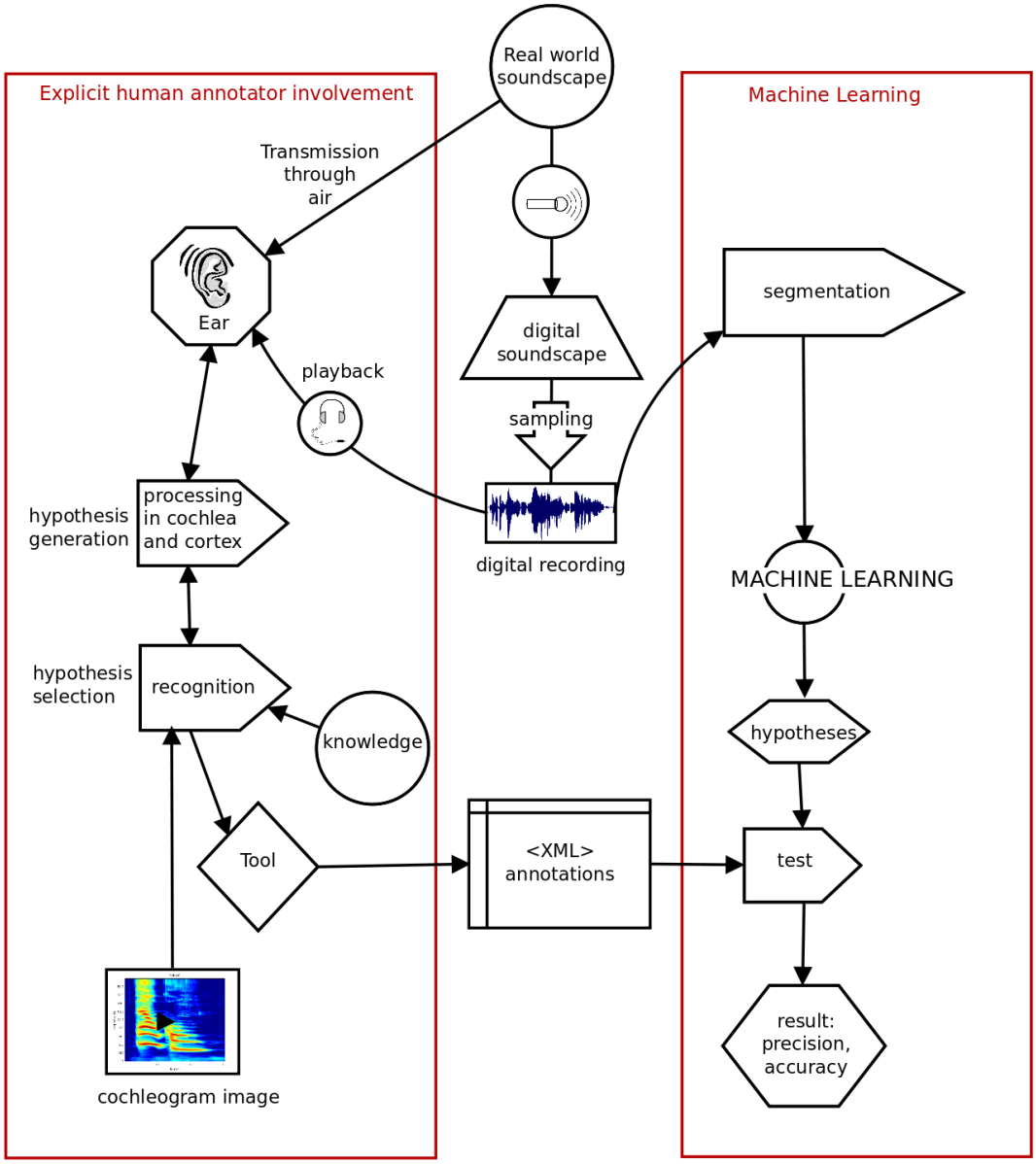

Figure 1.2: *Schema showing the relation between human annotation and sound recognition human soundscape annotations can be used to test a automatic recognizer. The role of the cochleogram image on the bottom right is introduced in chapter 4.*

applications are the following:

For *storage* and *retrieval*. Annotating a sound recording allows to search through the annotations to retrieve the requested (part of a) recording without listening to the sound. Also collections of sound files can be searched much more quickly by looking at the annotations instead of the sound data. An overview of techniques for audio retrieval is given in (Tzanetakis and Cook 2000a). Libraries have been using tagging methods for collections of audio and video recordings for a long time. In music retrieval,

social tagging is gaining a lot of interest, see (Levy and Sandler 2009) for an example.

To *train* sound recognition algorithms. Most machine learning paradigms require labeled examples to train the classifier. Annotated sound files can be used to train sound classifiers or sound segmentation algorithms.

To *test* sound recognition algorithms. Human annotations can be used as a baseline in determining the performance of an automatic sound recognizer. In a typical machine learning paradigm a data set (consisting of annotated recordings) might be divided in a training part and a test part. Figure 1.2 provides a schematic over view of a potential application for sound source annotations in an automatic sound recognition paradigm.

For soundscape *research*. Annotations provide an abstraction of the data contained in the sound recording, and this abstraction can allow researchers to easily extract segments of the data that are relevant to their research. Well-annotated audio recordings are also much more easy to inspect than non-annotated recordings. Several scientific disciplines can benefit from annotated soundscapes. Researchers interested in soundscape perception may use the annotation paradigm to collect people's perception of an environment; for example, in urban planning one may question how traffic noise from a busy road influences the inhabitant's perception of the soundscape in a nearby recreational area. In spoken language research the annotation to a sound recording allows the researcher to easily extract the parts of a recording that contain speech and hence are relevant, leaving out non-speech and therefore irrelevant parts.

Hearing aid validation (Grootel et al. 2009) mentions the (potential) application of annotation for the validation of electronic hearing aids: annotations from well-hearing people could serve as a ground thruth for validating the use of (a new type of) hearing aid in auditory impaired people.

1.0.2 Manual annotation: a time-consuming, tedious task

If a soundscape is recorded and is available for off line use, it is well possible to let listeners annotate that recording by manually entering annotations on a timeline, by letting them specify a textual description or class assignment for each sound event or sound source they recognize in the recording. This however takes a lot of time: anecdotal evidence indicates that subjects need around twice the length of the audio recording create an annotation with a moderate level of detail. In a study with a more or less comparable task by (Tzanetakis and Cook 2000b) subjects are reported to use around 13 minutes to annotate a 1 minute long recording; it took the participants much longer to fully annotate the recording than to listen to it. The manual annotation task is also considered tedious by subjects, as was found in an experiment with a early implementation of the annotation tool (Krijnders and Andringa 2009a).

This thesis seeks to develop an annotation method using a software tool that is both quick and accurate, and is not considered tedious by the user, as boredom or irritation may

decrease the quality of the resulting annotations.

There are, roughly speaking, three routes to take to achieve this goal:

1. Use *motivated, well trained annotators* and pay them to perform the annotation task both fast and precise. This is an expensive option because every hour of annotation will have to be paid out, and it is likely that there exists a ceiling effect for the learning curve for annotation, limiting the possibilities to speed up the process.

2. Dedicate the task completely to the computer: *automatic annotation*. Currently this is not possible for general sound recognition. In section 2.1.1 the current state of the art in sound recognition is discussed.

3. Let computer and annotator work together to achieve a good description: this can be viewed as either *machine learning with supervision* or *assisted annotation*, depending on the perspective that one takes (from the subject or from the computer). This approach is discussed in the Future Work section of this thesis.

4. *Embedded collection of tags*, for example through social tagging [1] use this technique) or games[2]. Currently no implementation of social or game tagging exists for annotating environmental sounds. A drawback of these techniques is that they may result in noisy tags.

In this project the first approach is taken: what happens when different annotators are asked to annotate a sound recording? How do they carry out their task? What labels do they choose for the sound sources they detect in the recording? These questions form the basis of this thesis.

1.1 Research questions

From the introduction above we arrive at the *research questions* for this master's project. The **main question** is formulated as follows:
How can a software tool assist the user in the task of annotating a real world soundscape recording?

Here we pose the following **subquestions**:

1. What is a 'good' annotation and which aspects of the annotation tasks influence the quality of the resulting annotations?

2. How does a subject perform the annotation task? Which aspects of the task can be supported by software?

[1] `http://www.Last.fm` and `http://www.pandora.com`

[2] Video tagging game *Waisda* uses this technique to collect annotations for Dutch television shows, see `http://blog.waisda.nl/`.

3. How can the annotation task be made less tedious? How can the time it takes to fully annotate a soundscape recording be shortened?

4. What is the role of auditory attention (see section 2.3.2 for a discussion of this phenomenon) in this domain? How can auditory attention be guided or supported by a software tool?

This master's thesis seeks to answer these question in detail. To find these answers, the project seeks to achieve the following **research objectives**:

1. Describe the current state of research in soundscape annotation. See chapter 2.

2. Implement a software tool for real-world soundscape annotation. See chapter 4.

3. Test this tool in an experiment and study the strategies and behavior of the participants in that experiment. See chapter 6 for this discussion.

In the next section relevant scientific literature for the topic of real-world soundscape annotation background will be discussed.

Chapter 2

Theoretical background

The first chapter of this thesis introduced the topic of the current project: *semantic annotation of soundscape recordings*. The introductory chapter explained that this task consists recognizing of sound sources in a recording, indicating the time region in which the sound source is present and selecting a semantic description for that sound source. This chapter provides a theoretical framework for the cognitive task of annotating a soundscape: in the view expanded in this thesis it is interpreting a recorded soundscape in an annotation task and constituting a set of sound source descriptions that describes the contents of the recording.

Before diving into the specifics of soundscape annotation and audition in general, the development of the field of sound recognition will be discussed, because this is most likely the area in which annotated soundscapes find their main application. Section 2.1 discusses the development of this field.

It is then important to define the 'input' of the annotation process: the kinds of soundscapes and recordings containing environmental sounds that are considered in this project. Therefore in section 2.2.3 a definition is provided for the 'stimuli' used in this project.

Listening to a recording of environmental sounds can be regarded as a special case of the general human ability to sense the world through the auditory system. Therefore, a more general account of listening is helpful to understand this task. A review from literature concerning general audition is provided in section 2.3 of this chapter. Literature reveals that the concept of *attention* is important in audition: attention can be seen as the *searchlight* of the auditory system (see subsection 2.3.2). This is not a unique feature: attention also plays an important role in other perceptual modalities. Because the importance of attentional processes was recognized earlier in vision, the discussion in this chapter first reviews the phenomenon for visual perception before reviewing similar processes in the auditory domain.

Attention processes need to function *upon* a representation of the input the system receives. It is proposed in section 2.3.3 that (theoretically) representing the perceptual input as *auditory gist* provides a reasonable framework for explaining attention and stimulus selection.

Section 2.3.4 hypothesizes that the building blocks of auditory perception can be described as *auditory objects* - these objects can provide a framework for the task of annotating a sound recording. This discussion leads to a 'recipe for an annotation tool': the last section shows how the theoretical topics discussed in this chapter lead to design choices for the an-

Figure 2.1: *The IBM Shoebox, a machine built in the sixties that performs arithmetic on spoken commands. Image ©IBM Corp.*

notation tool that this project seeks to develop.

It is important to recognize that In 'general' perception humans integrate information coming from all available sensory modalities to generate hypotheses about the state of their environment. Even when the primary source of information is the auditory system, assisting or conflicting sensory input from the other senses can be crucial to disambiguate complex streams of auditory input. In this thesis the focus is mainly on auditory perception, the current discussion will only touch multi-modal perception for a few times.

2.1 Automatic Environmental Sound Recognition: A field in development

The topic of automatic sound recognition was mentioned a few times already, this section will discuss this developing field in more detail. For the past decades, attempts to build automatic sound recognizers mainly aimed at transcribing speech and music automatically. One of the first scientific reports on automatic 'speech' recognition is the work of Davies and colleagues, which described a machine that could extract spoken digits from audio (Davies and Balashek 1952).
This reflects the early field's focus on developing machines that can perform typical office tasks automatically. This tendency can also be observed in the quest for creating an 'automatic typist', a dictation machine that transforms spoken sentences into text. This goal has been achieved: modern computers can be equipped with speech recognition software

that performs reasonably well under controlled circumstances. After an extensive training phase, typically taking more than an hour, a speech recognition application typically scores above 95 percent (on word basis) in recognizing spoken sentences correctly (Young 1996). This score is achieved by modeling speech as a Markov process. In this approach the system recognizes phonemes that are matched to a hypothesis of the further development of the spoken sentence.
However, these 'automatic typists' systems have major drawbacks: the training phase is too long for most users and the suggested sentences often need manual correction (and failing recognition often results in complete nonsense). Moreover, recognition only works reliably when the input is clean. This last problem illustrates a leading assumption in speech recognition: that the only speech is from the person dictating, that the microphone is placed close to the mouth of the speaker, that noise is limited to the minimum, and that there is a training phase in which the recognizer can adapt to a new speaker. When one of these assumption fails the recognition goes down rapidly, causing (potential) users to reject the technology.
Successful applications of speech recognition work either under clean conditions or in limited domains. In military applications, where it is crucial that soldiers keep their hands free for other tasks while providing input to electronic devices, spoken voice command recognition has reached serious applications and is used in practice. Other applications in which speech recognition is successful are home automation and automated telephone information systems.

Another example of 'automatic listening' that has gained attention in the past decades is *automatic music transcription*. Despite increasing effort put into this field, there still is no general, easy-to-use method for automatic music transcription (Klapuri 2004). Methods developed in this field again assume clean, structured input. Current techniques cannot handle mixed streams of audio; the field thus suffers from the same fundamental problem as described for speech recognition. This fundamental problem has to be solved for the field to succeed in its task.

Another field that connects auditory perception and sound technology is the field of electronic hearing aids. These devices allow the auditory impaired to take part in normal society. By capturing the sound that reaches the microphone and sending the amplified signal either through the ear canal, via the skull or even directly to the cochlea, an auditory impaired or even deaf person can gain relatively normal hearing capabilities again. Major achievements in this field are the development of seemingly invisible in-ear hearing aids and cochlear implants.
Both these applications require knowledge of the inner workings and physiology of the ear, especially the latter example where an electrode is implemented in the cochlea to compensate for an impaired mid-ear. The link between hearing aid technology and auditory perception research is however not as close as one would expect; the industry's focus is mainly on the sound technology, not on assisting auditory perception in a way that honours cognitive perception. Annotated sound recordings could provide a baseline for recognition of sounds in a person wearing an electronic hearing aid.

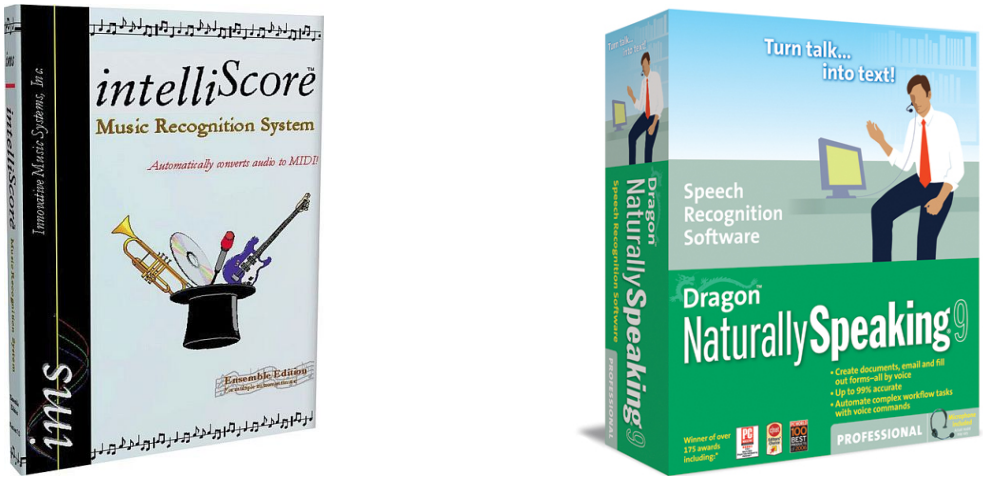

Figure 2.2: *Intelliscore is a software package that is able to transcribe recorded music to a MIDI file. Dragon NaturallySpeaking is a dictation tool for the personal computer.*

From the above discussion it becomes clear that the attempts to automatically recognize audio signals for long have focused on well-defined, very specific tasks and that no successful general approach to sound recognition has been found yet, nor do current approaches implement knowledge on the way humans interpret their auditory environment. Processing power and memory demands have increased over the years but probably are not the problem.
There seems to be a fundamental problem with computer audition that is limiting the breakthrough of automatic listening systems? An important cause of the inability of current systems to impress is the underlying assumption that the system should be able to function in a simplified version of the real task environment (Andringa 2010). The focus on a 'clean' signal in speech recognition is a clear example of this: for long the developers of these systems assumed that it is reasonable to ask the user to take care of the input, i.e. limit background noise, assure that the microphone is placed well, speak loud and clearly, etcetera. However, the true challenge for researchers in the field of automatic sound recognition is to build a system that, like humans, is able to function in an unstructured, real world environment. A system that stands this test has much more potential than do current end-user solutions.

A successful application of real world sound recognition is the aggression detection system presented in (Van Hengel and Andringa 2007). This processes street sounds to detect situations of (potential) aggressive situations. The challenge this system has overcome is to *ignore* most of the input; only a small percentage of the sounds that are analyze actually contain aggressive content. The principles that underly this system are described in section 2.1.1.

2.1.1 Towards robust, real-world automatic sound recognition

The previous paragraph concluded that despite the efforts in the past decades, in most cases automatic sound recognition algorithms currently only work well on narrowly and conveniently defined tasks, under laboratory circumstances and on simplified problems. How to build automatic sound recognition algorithms that are general, flexible and robust? (Andringa and Niessen 2006) recognizes this problem and describes a paradigm to develop open domain, real world automatic sound recognition methods.
The proposal of Andringa and colleagues is to start from the natural example: the natural human system that performs auditory perception in a flexible and reliable manner as inspiration for an algorithmic approach. Features used for recognition should be calculated with physical optimality in mind: for example, the time constant needed to create frame blocks as input for the recognizer are not compliant with natural systems. Furthermore physical realizability could be taken into account to prune the set of hypotheses about the world that the system generates. The authors furthermore argue for 'limited local complexity' when building a hierarchical recognition system: the different steps and corresponding layers should be guided by the nature of the input and underlying principles, not by mere design choices of the developers. Lastly, the most important principle for this thesis is mentioned: when testing an automatic sound recognizer, the input should be unconstrained and realistic. For decades systems have been build that function well in laboratory circumstances but fail in the real world; new methods need to be developed to tackle real-world problems.
Training and testing sound recognizers on real-world data is an important step in the development of robust and reliable systems; sound source annotations tailored for this purpose are crucial in this process. This project develops methods to obtain useful annotations for such real-world stimuli.

2.2 Real world sounds, soundscapes and recordings.

The previous section indicated a need for realistic stimuli to train automatic recognizers. What are then these *real world soundscape recordings*?
The notion of a 'good' annotation depends highly on the input (the soundscape recording) and the desired output of the annotation process (annotations tailored for a certain application). It is therefore important to define the characteristics of the soundscape recordings this projects seeks to annotate. This section discusses those characteristics, resulting in a definition for *real world soundscape*.

2.2.1 Gaver: the ecological account of auditory perception

This thesis focuses on *human* perception of environmental sounds. Gaver (1993) makes an important point that helps to understand how humans perceive their environment through auditory senses. He makes a distinction between *everyday listening* and *musical listening*; the former focuses on hearing (sonic) *events* in the world, while the latter relates to the sensory

qualities, such as the amplitude or harmonicity of the sound. Both listening modes refer to the *experience* a listener has when perceiving sound.
With his *everyday listening* account of auditory perception, Gaver argues for the development of an *ecological acoustics*, which entails a strong focus on the explaining human perception of complex auditory events (as opposed to primitive stimuli). This view is inspired by (Gibson 1986) who developed an ecological approach to perception in general. An important notion is that perception is direct and is about events, not about physical properties. Humans do not perceive (variant) physical properties, but instead process *invariant perceptual information*.
Where Gibson elaborated on this ecological approach for vision, Gaver was the first to constitute a framework for understanding hearing and listening from an ecological perspective.

An important observation in Gaver's approach to environmental listening is that sounds are always produced by interacting materials. Sounds reveal to the listener attributes of the objects involved in the event; in the article the different physical interactions and the resulting sonic events are described extensively. Sounds also convey information about the environment as they are shaped by that environment when surfaces reflect the sound or when the air transports and shapes it (for example in the Doppler effect).
Gaver concludes from his own experiments that peoples' judgments of sounds correspond well to the physical accounts of acoustic events, and he argues that the combining peoples' reports with a physical account may reveal categories for sound perception. Based on this combined information Gaver builds a 'map of everyday sounds' that provides a hierarchical structure in which distinction between vibrating solids, liquids and aerodynamics forms a the basis.
Gaver's attribution to the understanding In his 1993 companion article *How do we hear in the world? Explorations in Ecological Acoustics* (Gaver 1993) focuses on *how* people listen: what algorithms and principles do humans exploit to analyze the acoustic input. This article is less important for this thesis, as it is aimed at other researchers implementing the algorithms and strategies into their sound recognition applications or model of audition.

How then to collect complex, real world audio recordings that can be used to study human audition? An important aspect of a sonic environment to consider is the amount of *control* the researcher can impose on the acoustic events that end up in the resulting recording; this is discussed in the next section.

2.2.2 Control in sonic environments

Sonic environments typically contain multitude of sound sources that vary in prominence. For recordings made in a *controlled environment* the number of sound sources is likely to be limited. As an example, let us consider a soundscape captured in an office environment: this recording may contain just a few prominent sound sources, such as the constant hum of an air conditioning system together with sounds of a worker using a computer and some occasional speech sounds. In an even more controlled environment, the researcher may assure

that in each recording there is only one sound source present, and that the beginning and ending of the sound are captured and clearly recognizable.
In a less controlled setting, such as a typical urban soundscape of a busy street, one can expect a mixture of sound sources to be captured; some sounds were already present when the recording started and some acoustic events may still continue when the recording ends. The presence of nearby traffic, multiple people talking, or the distant sound of the clang of metal coming from a construction site may result in a cacophony of sounds. Such a complex sonic environment makes it difficult for a human listener to recognize the events that occured during the recording, and moreover incomplete, mixed or masked sonic events make it hard to discriminate between sound sources.
Imposing control on the environment is one way of influencing the characteristics of a soundscape recording. Another way is to apply *acoustic filtering* to enhance the recording. As described above, we require sound recognition methods to be robust against noise, transmission effects and masking. Therefore in this project the minimum of filtering and noise reduction was applied when recording soundscapes. We do however allow ourselves to protect the microphone against direct wind influences that distort the recordings. It is reasonable to do so because the human anatomy also prevents the noise created by wind to distort auditory perception. For a more advanced recording method we refer to (Grootel et al. 2009) who uses an 'artificial head' to mimic transmission effects caused by the human anatomy.
A clear dividing line between controlled and uncontrolled environments cannot be drawn; it is better to define a continuum here, with the situation of a limited number of sound sources and completely controlled lab conditions on one side, and completely uncontrolled and mixed sonic environments on the other side. Sound recognition research for long concentrated on the former half of this spectrum (the 'easy' task), but now needs to focus on the 'hard' task in the latter part of this spectrum to overcome fundamental problems with the current approach. This project therefore seeks to provide annotations for recordings in the 'hard' area of the spectrum: real-world, uncontrolled environments.

2.2.3 Defining environmental sounds

In this project we study the perception of the kind of sounds that can be found in any environment where humans may reside, and we describe these sounds with the term *'environmental sounds'*. This term however lacks a common understanding: a debate amond sound researchers is ongoing on what exactly can be understood by this term. The same class of sounds is sometimes described as *everyday sounds*, for example in the work of Grootel et al. (2009). Gygi and Shafiro (2010) proposes a definition of environmental sounds - the article describes the creation of database that provides such sound recordings. Pointing at Gaver (as discussed above), Gygi states that determining the *source* of the sound is the goal of what he calls *everyday listening*. Therefore in Gygi's database *'the atomic, basic level entry for the present database will be the source of the sound'*. Gygi and colleagues do not state that only isolated sound sources are allowed in the database, but it does imply the exact location in time of a sound event needs not to be stored. In the view expanded in this thesis, the ap-

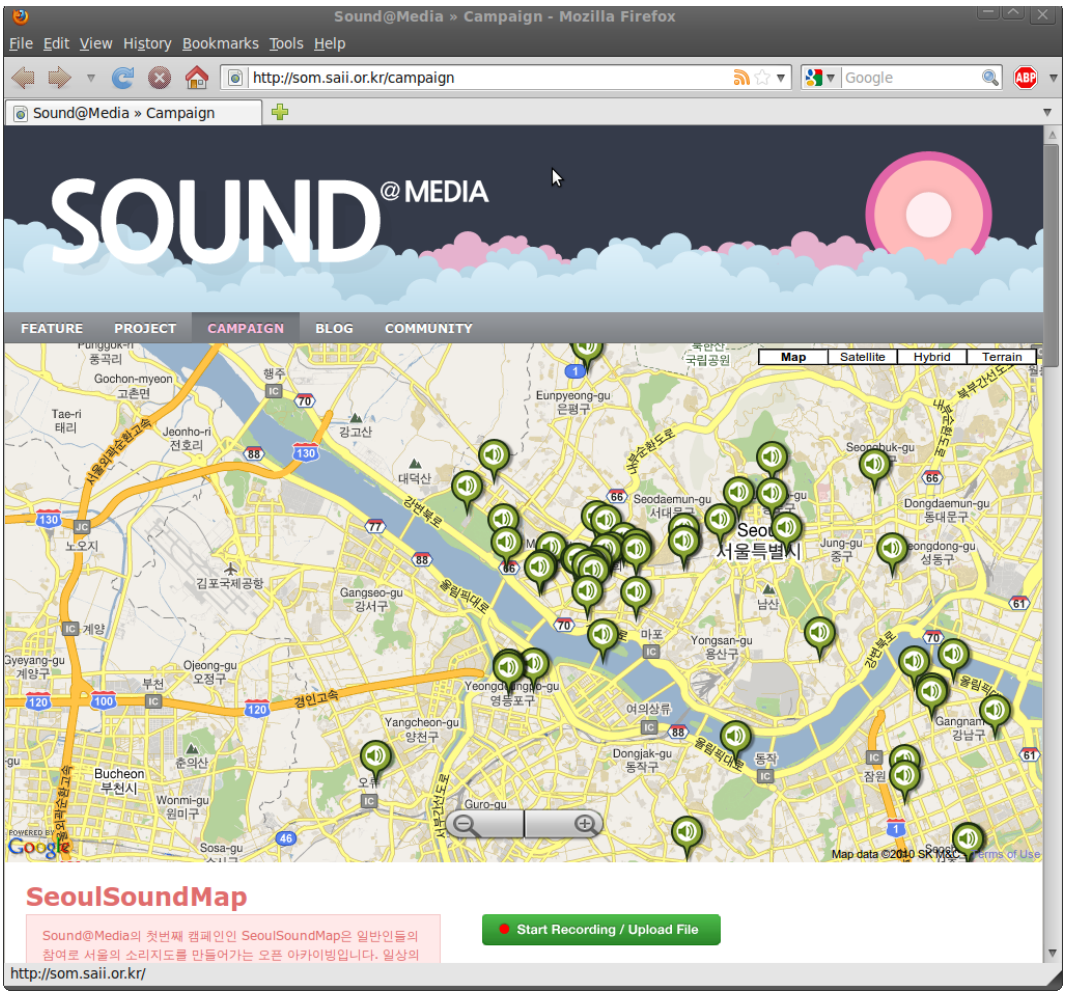

Figure 2.3: *Soundscape recordings can also be found outside research. This Korean website allows users to upload their own recordings of soundscapes that they find defining for their experience of the city of Seoul. Visitors can click on the map to get an impression of what Seoul sounds like.*

proach Gygi takes limits the possibilities for applying this database as input for an automatic sound recognizer. In this thesis a different approach is taken in which the (time-)location of a source within a recording *is* important. If a learning algorithm is not provided with data on segmentation (or: location in time) of different sound sources within the recording, this makes the task unnecessarily difficult. In this thesis it is assumed that providing detailed descriptions of sound sources contained in the recording, combined with timing data, is fair and is necessary if the annotations are used to train automatic sound recognition algorithm upon.

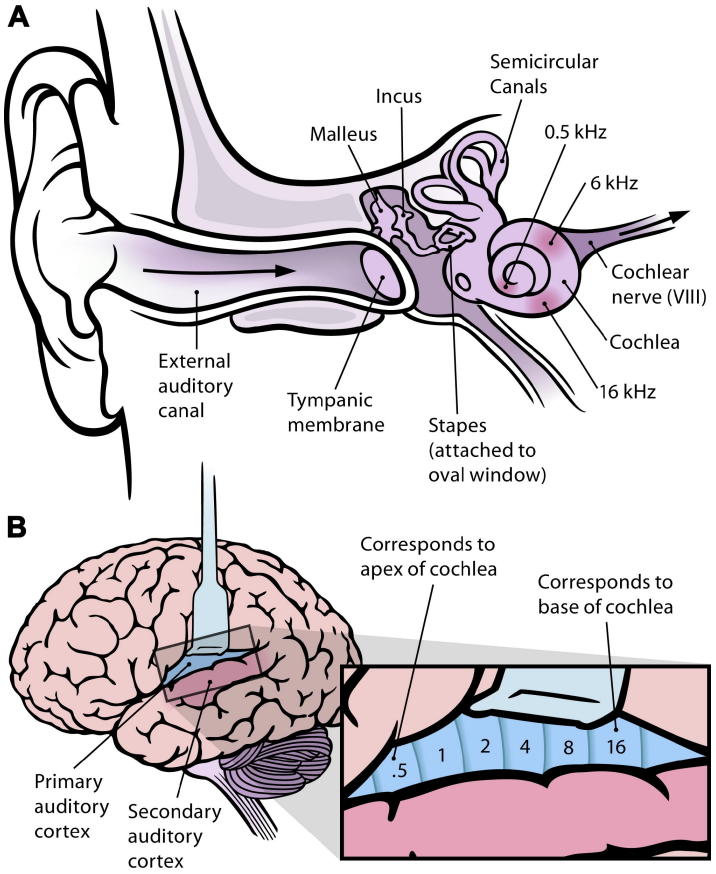

Figure 2.4: *Frequency Coding in the Human Ear and Cortex. From: (Chitka 2005)*

A definition for environmental sounds

The discussion of different visions on environmental sounds (see above) illustrates the need to define the stimuli for this project. Therefore we constitute our own definition here. This definition should include the notion of control as it was previously discussed, and should underline that we focus on sounds with real-world complexity and without smoothing or filtering applied.
For this project we define real world environmental sounds as: *Soundscapes recorded in a real-world, uncontrolled setting, without intervention other than needed to establish the soundscape recording.*

2.3 Audition: Hearing, listening and auditory attention

Now that a definition of real world environmental sounds is established, one can question how humans perceive sounds in a real-world setting. We first look into the general phenomenon of audition before turning to the specifics of listening for annotation.

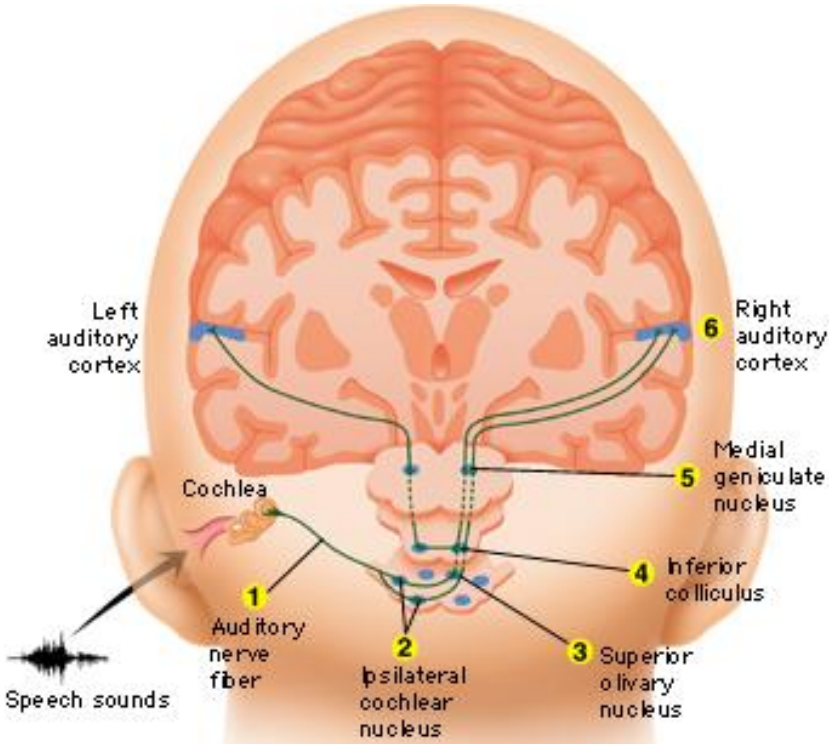

Figure 2.5: *Human auditory pathway. The outer, middl and inner ear are shown in figure A; figure B shows the auditory cortex. From* `http://brainconnection.positscience.com/topics/?main=anat/auditory-anat2`

The human auditory system is a highly sensitive, highly adaptive multi-purpose system that is capable of segmenting and recognizing complex and mixed 'streams' of acoustic information. The 'circuitry' involved in perceiving sounds is distributed over different organs. The outer ear is shaped so that it can capture sounds coming from the direction that the listener attends to; it leads the incoming sound through the external auditory canal to the timpanic membrane that vibrates with the sound. The timpanic bones (*malleus, incus*) then transfer these vibrations to the snail shaped *cochlea*, which can be seen as a tightly rolled up sensor array. The relation between place and selective frequency of the 'sensors', the hair cells, is a logarithmic one. The hair cells in the fluid-filled cochlea respond to different frequencies; humans are typically able to hear in the range between 20 Hz and 20 kHz. Low frequencies are captured at the base of the cochlea, higher frequencies are captured by hair cells further along the cochlea. Information from each hair cell is transfered through the auditory nerve through the brain stem to the left and right auditory cortex, where each frequency-responsive area of the cochlea maps to a cortical area that responds to activity for that region of the frequency plane. From there activation may spread through other parts of the cortex for further processing. 2.4 shows the most important structures; figure 2.5 shows how nerve cells pass the brain stem to the cortical areas.

However, not all incoming acoustic information is processed to the same level of detail:

situational and task-dependent factors influence the level of processing of a stimulus. This selective process is called *attention* and plays an important role in perception, not only in audition. Section 2.3.2 will explore the phenomenon of attention further.

Humans have the ability to recognize (reconstruct, in a sense) the nature of environment from the soundscape it produces, segmenting the input into 'streams' corresponding to separate sound sources. This process has been studied for about two decades under the term *auditory scene analysis*; subsection 2.3.1 covers this approach. Theories for ASA however has not lead to a comprehensive framework that explains how humans through audition are able to comprehend sound sources under complex circumstances, nor have other theories of human audition (Shinn-Cunningham 2008). However, recent developments in auditory perception research indicate that attention is a key concept that might explain the human ability to give meaning to complex, mixed auditory scenes; subsection 2.3.2 discusses this topic.

A theory that promises to be helpful in explaining auditory stimulus selection is the concept of *gist*. Theories that take this concept into account generally contrast the classical paradigm that auditory perception is a staged, hierarchical process. When accounting for top-down attention in human audition, a description in terms of related, parallel processes that influence each other seems much more accurate. Subsection 2.3.3 discusses *gist perception* in detail, both in the visual and auditory domain.

The main topic of this thesis is soundscape *annotation*; therefore, the aforementioned issues are connected to the soundscape annotation paradigm in section 2.4.

2.3.1 Auditory Scene Analysis

Psychologist Albert Bregman has described the human ability to organize sound into perceptually meaningful elements with the term *auditory scene analysis* (Bregman 1990). In the view expanded by Bregman, the incoming stimuli are organized into *streams* that the cognitive system can attend to. He argues that segments are formed from the fuzzy stream of acoustic information that is captured in the inner ear. These segments are then either integrated into one auditory stream, or segregated into different streams. Grouping can occur over time; related segments that occur in sequence can be grouped as one stream, according to *gestalt* principles. Grouping of co-occurring auditory events can also occur.
The streams that are formed in this process are thought to be related to events in the real world. In this view the constitution of auditory streams can be seen as the reconstruction of the acoustically relevant events from the acoustic environment.
This acoustic environment is described as the 'auditory scene'. This term refers to more or less the same concept as the term 'soundscape', with the difference that a soundscape is by definition the *perception* of the acoustic environment shaped as it is interpreted by the listener.
In practice ASA has mainly focused on impoverished stimuli such as tones, noises and

pulses; the theory lacks the explanatory power to account for the perception of real-world stimuli.

CASA: A computational approach

There have been attempts to transfer the previously describe human scene analysis abilities to a computational approach (Wang and Brown 2006) to automatically interpret the sonic environment. Brown and Cooke (1994) used the ASA paradigm to implement an algorithm that performs speech segregation. This computational approach models to some extend the periphery and early auditory pathways. Different stages model the process of feature extraction, the formation of auditory segments and the grouping or segregation of different streams.

Later work on this topic also has a strong focus on segregating speech from 'background' sounds. Despite early attempts to integrate speech perception and general audition theories (Cooke 1996) into ASA, the theory is not yet capable of providing a general account of audition. A central concept that is missing from the theory is *attention*. The next section will elaborate further on this important concept.

2.3.2 Attention: controlling the flood of perceptual input

To understand and evaluate the theory of selection of auditory 'streams' as proposed by those advocating ASA (as discussed in the previous subsection), it is helpful to take one step back and assess the general concept of attention. From this general discussion the focus will return to auditory stimulus selection.
Recent theories relate the ability to switch between 'streams' of auditory information to mechanisms of attention that guides perception. The selection and enhancement of perception goes mostly unnoticed, as philosopher Daniel Dennet points out:

> *The world provides an inexhaustible deluge of information bombarding our senses, and when we concentrate on how much is coming in, or continuously available, we often succumb to the illusion that it all must be used, all the time. But our capacities to use information, and our epistemic appetites, are limited. If our brains can just satisfy all our particular epistemic hungers as they arise, we will never find grounds for complaint. We will never be able to tell, in fact, that our brains are provisioning us with less than everything that is available in the world.*
>
> - Daniel Dennett in *Consciousness Explained (1991)*

As Dennett puts it, the sensory organs the brain is flooded with information from and about the environment. The capacity of the brain and nervous system to capture and process this information is limited, therefore this constant stream of incoming stimuli information needs to be filtered and abstracted. The mechanism that guides the selection of stimuli that need

to be processed further is called attention.

But what is the essence of attention? Early psychologist William James described it as follows in his 1890 book *The principles of psychology* (re-published as James et al. (1981)):

> *Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German.*

In James' description of attention two important components can be distinguished: an active *focus* on *objects*, and a more passive *taking possession* of what is attended to. These two components are currently distinguished as signal-driven and knowledge-driven processes of attention:

Bottom-up, signal driven attention is evaluative in nature, leaving unimportant stimuli unattended and elevating salient, relevant stimuli in the stream of information for further consciousness processing.

Top-down, knowledge driven attention is an open process that structures the perceptual input based on context and memory. This form of attentional selection elevates regions (by focal attention), features (by feature-based attention) or on objects (by object-based attention) from the scene, at the cost of ignoring all other stimuli.

Context is important here: the situation, the physical surroundings and the expectations of the perceiver may raise expectations (hypotheses, one can say) for which 'evidence' is sought in the low-level signal representations.
The knowledge 'objects' that are formed in this structuring process are then available as input for reasoning about the state of the environment. These concepts will be discussed for auditory perception later in this thesis. First, the phenomenon of attention in the auditory and visual domain is discussed below.
The concept of *consciousness* is closely related to attention; James already linked the two phenomena in the quote above. It should however be noted that both concepts still lack a clear definition; consciousness may refer to subjective experience, awareness, a person experiencing a 'self' or may refer to the executive control system of the mind. For the current discussion the last interpretation of consciousness is adopted. There is an ongoing debate on the relation between and dissociation of these two phenomena; some argue that attention is necessary for conscious perception (Dehaene et al. 2006), while others argue that both may occur without the other (Koch and Tsuchiya 2007).
Attention is also intimately linked to the creation and storage of new memories: it is argued that attentional processes serve as a 'gatekeeper' mechanism for stimuli to reach awareness and to be stored in memory (Koch and Tsuchiya 2007); without attention a stimulus cannot reach declarative memory.

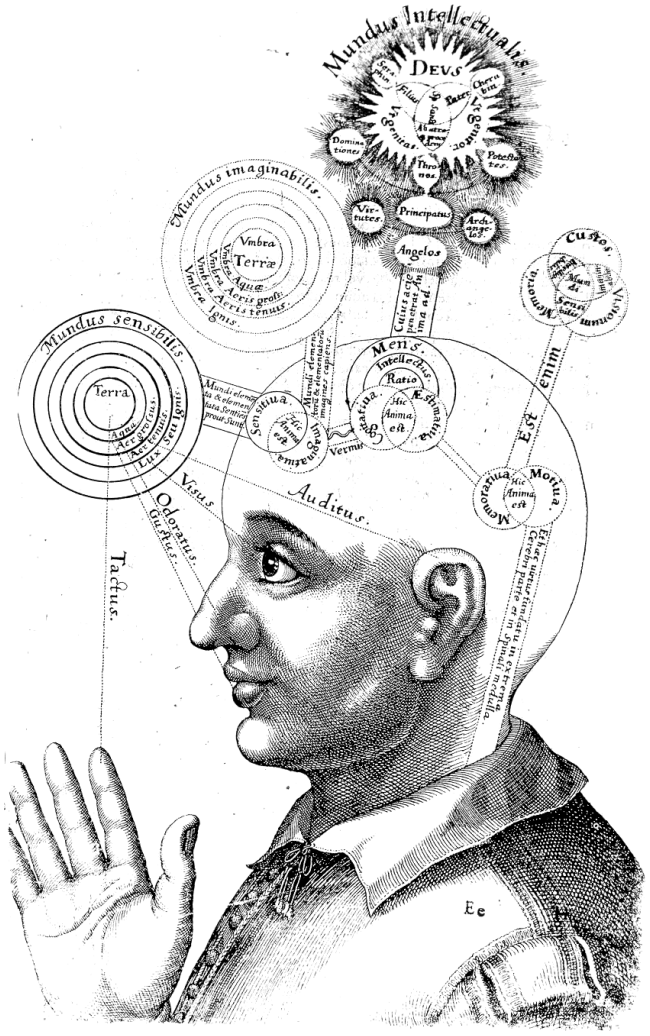

Figure 2.6: *Consciousness as it was seen in the seventeenth century by Robert Fudd. Original source: Utriusque cosmi maioris scilicet et minoris [...] historia, tomus II (1619), tractatus I, sectio I, liber X, De triplici animae in corpore visione*

The remainder of this subsection will discuss research into attention phenomena. The following route is taken: first the work of Cherry is presented, which revealed attentional influences in audition. Before discussing auditory attention further, we look into attentional phenomena (and possible undesired effects) in vision. A comparison between attentional in auditory and visual attention is made. Before turning to the concept of *gist*, another cause of stimulus omission is presented: auditory masking.

Attention research: Cherry and dichotic listening

In a classical experiment Cherry (1953) presented subjects with two speech signals, each speech signal presented to an ear ; see figure 2.7. When asked to attend to one speech signal and listen to (comprehend) the message that this voice brings, the subjects showed unable to report what the other voice was talking about. Not all information from the unattended ear

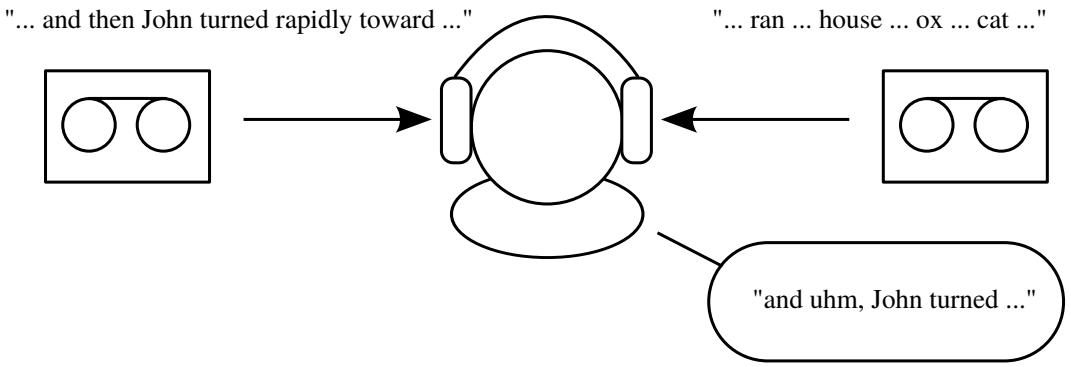

Figure 2.7: *Drawing of the dichotic listening task. Two audio signals are presented to each ear of the participant and he/she is then asked to attend to one of the signals. Afterwards the subject is asked what he/she knows about the unattended signal.*

was ignored: basic aspects of the unattended speech signal were correctly reported by most subjects, such as the gender of the speaker. Cherry found that when salient features were used in the unattended speech signal, such as the first name of the subject, attention could be shifted to that stream of information very rapidly and would allow the participant to report the stimulus. This indicates that not all information was ignored from the unattended ear; some processing must occur to allow the subjects to recognize the salient words.
Where Cherry was a pioneer in discovering the *cocktail party effect*, Broadbent (1958) was the first to formulated an extensive theory on selective attention, the *filter theory*. For the theory Broadbent was inspired by the computer, which he used as an analogy in explaining the limited processing capacity of the brain; all possible features are extracted from the input, and then filtered for the part of the input that is attended. This 'late selection' theory was contrasted by 'early' selection theories that emerged later; the debate over these theories has still not been resolved. A further discussion of the different theories on attention can be found in (Driver 2001).

Attention: Common mechanisms for the auditory and visual domain?

The phenomenon of attention has been studied extensively for the visual domain, however for the auditory domain it has only recently been explored, due to both technical and conceptual difficulties (Scott 2005). An extensive review of current neurobiological research in auditory attention is given in (Fritz et al. 2007). In this discussion it is concluded that the phenomenon of auditory attention is produced by a rich network of interconnected cortical and subcortical structures. This network is highly flexible and adaptive; depending on the task, relevant subnetworks can be invoked to enhance the input. Research indicates that top-down influences can also influence the shape of the receptive fields in the auditory system, thereby influencing perception on the lowest level possible, in the cochlea (Giard et al. 1994).

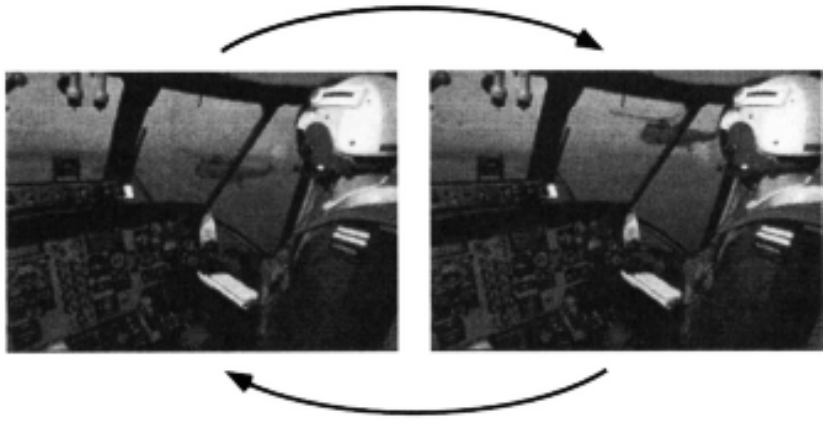

Figure 2.8: *An example of a scene in which the spatial location of a stimulus is changed. When showing these two images with a 'flicker' in between, most viewers do not see the change in location of the helicopter that can be seen through the windscreen. From (Rensink et al. 1997).*

Visual attention 'deficit': Change blindness and inattentional blindness

The selective nature of the mechanism can cause the brain to 'miss' events or changes in the state of the environment. These omissions demonstrate important principles that underly the phenomenon of attention.
An example of such an omission of a stimulus in the top-down attention process can be observed in a phenomenon that is called *change blindness*. This is the case when a change in a non-attended stimulus is not consciously perceived. In vision, this inability to become conscious of (possibly important) changes in a scene has been investigated extensively, both in controlled lab conditions and in more naturalistic scenes. The effect is even stronger when the change is unexpected, for example when in two (otherwise identical) photographs are manipulated so that the heads two persons are exchanged and these two versions of the picture are shown consecutively.

A review of this striking phenomenon in visual perception can be found in (Simons and Rensink 2005). This review argues that there is a close link between this change perception, attention and memory. Attention is key to consciousness perception, and the limited attention resources that are available to the sensing brain prevent it from 'seeing everything that is visible'. The explanation given by Simons and colleagues is that when not attending to the aspect of a scene that changes (or: the part of the perceptual input space where the event happens), it is likely that this aspect of the perceived scene does not reach awareness and memory. Therefore the 'new' scene (after the event occurred) cannot be compared to the previous version, causing the change to go unnoticed. An example of a hard-to-detect change in scene can be found in figure 2.8[1].

[1]For more examples, see `http://nivea.psycho.univ-paris5.fr/#CB`

Figure 2.9: *Screenshot from a video recording of one of the scenarios from the 'Gorilla experiment': while focussing on the actors dressed in white, the black-suited gorilla was complete missed by most subjects. Image from (Simons and Chabris 1999).*

Another example from the visual domain of stimulus omission through a top-down attentional process is *in-attentional blindness*. This is the phenomenon that a stimulus that otherwise would be perceived may even be completely missed when attention is directed to another stimulus.
A well-known example is the 'Gorilla test' described in (Mack 2003) where subjects were told to observe a group of people dressed in black and white t-shirts playing a ballgame. Their task was to count the number of passings by the white team, while ignoring the ball passings of the black team. While attention was focused on the ball, a person in a (black) gorilla suit crossed the group of players. Upon questioning the subjects afterwards, most subjects appeared to have completely missed the gorilla, whereas this normally would have caught their attention, see figure 2.9. This is an example of how focusing the attention on part of the perceptual input space may cause important and salient visual events elsewhere in the input space to be missed completely. A more formal approach and demonstration of in-attentional blindness can be found in (Mack and Rock 1998).
Both these phenomena may be explained as inherent properties of an otherwise very accurate attention system, but the most compelling aspect of these forms of attention deficits is that people tend to be blind for their own blind spots in perception. Most of the subjects in the 'Gorilla test' were absolutely sure that there was no monkey present in the video, but immediately admitted that they were wrong when reviewing the footage. This shows how both the percept and the decision in involuntary omissions of stimuli causes these stimuli to not reach conscious awareness at all.

Perceptual deafness: Stimulus omissions in audition

From the above discussion it has become clear that in vision omission effects can be observed; can these principles also be shown for audition? If so, does this mean that auditory and visual perception have common underlying principles and even share neural structures?

Support for *change deafness*

Literature does report observations of omission phenomena in in audition that can be related to the similar observations in vision. (Vitevitch 2003) presents an experiment that points in the direction of the existence of what might be called *change deafness*. In the experiment that Vitevitch et al. describe subjects were asked to listen to a speaker reading out a list of words. Halfway through this list the speaker was replaced by second voice. The researchers observed that around 40 percent of the subjects missed this important change in the stimulus and hypothesize that this was due to a strong attentional focus on the task (memorizing the words). While this effect showed to be stronger when the change in the presented stimulus was introduced after a one-minute break, it was also observed when this break was much shorter. Even when speakers were mixed across words (and hence many changes occurred and went unnoticed) the effect could be demonstrated.
(Eramudugolla et al. 2005) investigates this phenomenon for a more complex auditory scene, namely a musical setting where one instrument is added or deleted from an acoustic scene. The results of the described experiment indicate that when the attention is directed to this instrument (by showing a word on a screen to prime the subject, for example 'piano') the inability to become conscious of the change in the scene disappears almost completely. The effect was also observed when the spatial location of an instrument was changed; in the experiment this change was modeled by shifting the stimulus to a different speaker in a multi-channel audio playback system. The assumption that a 500ms period of white noise is needed to 'mask' this change was challenged by (Pavani and Turatto 2008), and disconfirmed in that publication. Pavani and Turatto hypothesize that the inability to detect changes in auditory scenes is caused by limitations of the auditory short term memory, not by limitations perceiving auditory transients. The nature of these limitations remains debated: McAnally et al. (2010) argues that generally no explicit comparison of objects occurs in change detection, but however that it is well possible that an explicit comparison process is invoked. For this to happen, enough information needs to be available the process is probably limited by the system's capacity to parse an auditory scene.

Differences between change deafness and blindness

The current discussion provides evidence for the existence of 'change deafness' and similarity between the two modalities, there are however differences between vision and audition: Demany et al. (2010) reports special properties of the auditory counterpart. The authors argue that their results indicate that auditory memory is stronger over relatively long gaps and

is in some cases stronger than visual memory. According to the view expanded in the article this is due to largely automatic and specialized processes that provoke the detection of small changes in the auditory input. More research is needed to explore the common nature of omission phenomena in both modalities.

Support for *in-attentional deafness*

Literature supports a form of 'change deafness', as discussed above, but is there also evidence for *in-attentional deafness*, the counterpart of in-attentional blindness? It seems that this is indeed the case. An intriguing demonstration was given by famous violin player Joshua Bell who played Bach's *'Chaconne'* in a subway station - from the more than 1000 people passing, only 7 stopped to listen to Bell and his Stradivarius [2]. The rest of the people traversing the station seemed to fail to detect his prescence.
This may also be seen as a failure to recognize the exceptional quality of Bell and his instrument. More firm support for the pnomenon of inattentional deafness can also be found in literature, the experiment by Cherry (discussed in the opening of this chapter) can be seen as one of the first well-documented references to its existence. A more recent and structural report can be found in (Sinnett et al. 2006) which presents experimental evidence concerning in-attentional blindness (in both the auditory and visual domain) and describes an experiment that demonstrates cross-modal attention effects. In this experiment participants were asked to monitor a rapid stream of pictures or sounds, while concurrent auditory and stimuli (spoken and written words) were presented. In-attentional blindness was reported for both auditory and visual stimuli. The 'blindness' effect appeared to be weaker when attention was divided over the two sensory modalities. The hypothesis of the researchers is that *' ... when attentional resources are placed on an auditory stream, unattended auditory stimuli of a separate stream may not be consciously processed.'* More evidence indicating a cross-modal nature for change 'blindness' is presented in (Auvray et al. 2007).
An alternative explanation for the inability to become aware of changes states that the problem is not in the recognition phase but in memorizing the unexpected stimulus: by the time participants have to report if they saw or heard an unexpected stimulus they might simply have forgotten that they did. This *'inattentional amnesia'* theory is not very powerfull as it fails to explain the observation that the effect also occurs in stimuli that one would expect the participants to remember, for example the view of a person in a gorilla suit passing the scene in the work of Simons and Chabris.

[2] See `http://www.washingtonpost.com/wp-dyn/content/article/2007/04/04/AR2007040401721.html` for a report of the experiment.

The role of the temporal dimension in audition

(Shamma 2001) describes from a neurobiological point of view how auditory perception works over time; the article discusses different models that may explain how cochlear input is processed over time and argues for a unified network that computes various pereceptual sensory input, including audition. In this view the basilar membrane is thought to be responsible for the transformation of temporal cochlear excitation patterns into cues that can be processed by the same neural structures as for visual perception.

(Cooke and Ellis 2001) beschrijft dit ook.

Auditory Masking

Attentional omission effects are not the only phenomena that cause a person to miss an auditory stimulus; another important source for omissions is that of *auditory masking*, or *acoustic masking*. This the phenomenon that the occurrence of another stimulus (the masker) can influence the perception of a stimulus (this signal). Different forms of auditory masking can be distinguished. Masking can occur simultaneously or non-simultaneously: either the masker occurs together with the signal or the masker occurs before or after the signal. In both cases the masker can make the signal less audible or completely inaudible. The strength of the this masking phenomenon depends on frequency of the masker: when masker and signal property frequency are close together, the two sounds can be perceived as one. This effect is stronger for high frequencies than it is for lower frequencies, an observation that is called 'upward spread of masking'. This upward spread is due to the size of the filter that is applied in to the incoming sound in the cochlea.

Masking can occur when the masker and stimulus are presented on one ear only (ipsilateral), but can also occur when masker and stimulus are presented either one ear (contralateral). This effect is due to interactions in the central nervous system. One of these interactions is called *energetic masking* (Shinn-Cunningham 2008). In this case the system is too 'busy' responding to a co-occurring stimulus to respond to an auditory event. Other forms of masking are often called *informational masking*; Shinn-Cunningham (2008) argues that these masking phenomena are due to failures of object formation. This topic will be discussed below.

informational masking: `http://scitation.aip.org/journals/doc/JASMAN-ft/vol_113/iss_6/2984_1.html`

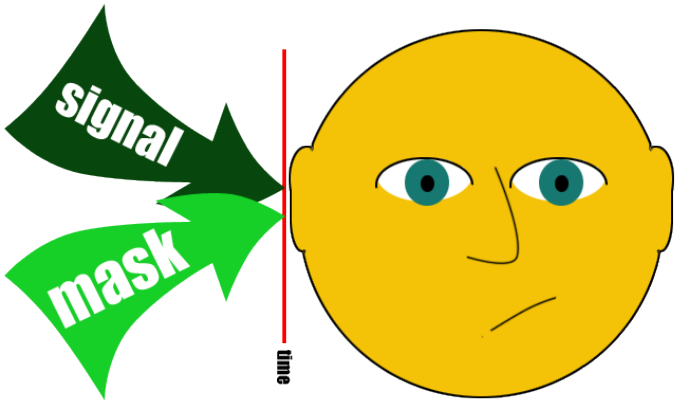

Figure 2.10: *Auditory masking: the co-occurence of the masking signal influences the perception of tne target signal.*

2.3.3 Gist perception

The previous section discussed the phenomenon of attention in both the visual and the auditory domain. This makes clear that attention plays a major role in perception: attentional stimulus selection appears to be very important in the process of shaping a perceptual experience. From this discussion the question might rise what exactly is the information that attention processes work upon. The concept of *gist* might explain important characteristics of visual and auditory perception. In audition, gist theory conflicts with the view that was expressed in section 2.3.1: the view that the human brain hierarchically processes of all auditive input into streams is probably incorrect. Each stream would have to be processed to a certain detail to allow attention to select which stream needs to reach awareness, which would pose a high computational demand on the brain. Introducing auditory gist provides an explanation that fits experimental data better, and furthermore explains the interplay between bottom-up, signal driven recognition and top-down, knowledge driven influences.
The concept of gist was originally formulated for vision, therefore this modality is discussed first.

Gist in vision

For the visual domain (Oliva 2005) introduced the concept of the *gist* of a perceptual scene to explain why humans are able to interpret a visual scene very quickly. Oliva describes the gist as:

> *'... a representation that includes all levels of processing, from low-level features (...) to intermediate image properties (...) and high-level information(...).'*

In vision one can intuitively envision how such a representation would look like: when blurring the visual scene, only a raw representation remains and unimportant details are

FIGURE 41.3 Illustration of a scene gist representation that conserves sufficient structural cues to infer the probable category of the scene. This global scene representation is used to determine spatial envelope properties of a scene. The information preserved by this global representation is illustrated on the right-hand image: it represents a sketch version of the original scene, computed by coercing noise images to have the same global features as the left scene (Torralba and Oliva, 2002). The scene sketch corresponds to an "unbound" spatial layout representation of contours, texture density and colors in the original scene picture.

Figure 2.11: *From (Oliva 2005). - Description -*

lost, but recognition of the most important aspects of the scene is still very well possible. For an example, see figure 2.11.

This representation is constituted within 100 milliseconds after the appearance of the visual stimulus, and its purpose is to allow fast image identification. Oliva distinguishes the earlier *perceptual gist*, which originates from the percept and holds a structural representation of the scene, and *conceptual gist*, which contains (semantic) information on the concepts that are recognized in the scene after processing. The latter can lead to a verbal description of the scene that can be reported by the subject.
Once a conceptual representation is established, it is candidate for storage in long-term memory. In vision, the spatial properties of the scene form an important aspect of the representation. The gist also contains rudimentary statistical information, including an estimation of the number of objects in the scene. The *spatial envelope theory* states that the gist is built quickly directly from low-level features without applying a hierarchical processing scheme. It is important to note that this theory contrasts with classical theories on visual perception. These theories are well outside the scope of the current discussion, but is is likely that gist theory will greatly influence the common view on visual perception.

Gist in the auditory domain

From the above discussion of gist in the visual domain it becomes clear that gist perception can be seen as performing a very quick 'scan' that provides a basis for further, attentionally influenced processing. Top-down, knowledge driven information then influences which part of the input is processed further.
This concept can also be applied to audition, where one can postulate a preprocessed form

of the auditory information that allows attention processes to swiftly select to what part of the auditory input attention should be directed.
In principle this view conflicts with conventional theories that assume hierarchic processing of (all) low-level information to higher level auditory *streams* as described in section 2.3.1, but is does comply with the observation that humans are able witch attention within ?? ms between auditory streams. The postulated processing of all low-level information requires a lot of computation that probably would take more time than experimental data indicates .

The role of gist in hearing and listening

Harding et al. (2007) reviews the concept of auditory gist and discusses the existence of a minimal representation that contains just enough information to provide an overview of the acoustic scene. The review contrasts theories that describe attention as auditory stream selection and proposes a theory that implements the concept of gist.

An important insight of Harding et al. is distinction between *'audition at a glance'* and *'audition with scrutinity'*, where the former refers to a process resulting in a gist-like representation and the latter relates to exploring the input in a top-down fashion with the gist as a guideline. This distinction comes at hand when dissociating between *hearing*, which can be seen as the 'bottom-up gist producing stage' in the words of Harding et al, and *listening*, which results in an overview of the whole scene by top-down activation of relevant concepts. Only in the listening-stage attentive, detailistic analysis can take place and higher order concepts are enabled to reach consciousness. In conscious, everyday auditory perception these processes constantly work together to present the brain with a stuctured interpretation of the stream of sound. Andringa (2010) refers to the resulting dynamics as the 'flow of consciousness; figure 2.12 schematically shows how this results in a person perceiving (here:) the sound of footsteps and a voice greeting the listener.

Computational gist features: An example

An implementation inspired by the concept of auditory gist is described in (Kalinli and Narayanan 2009) The proposed algorithm uses both top-down (task-dependent) en bottom-up information to determine the amount of stress and prominence for regions of speech recordings. A cochlear model is used to model the early auditory system, from the resulting auditory spectrum five features are extracted: *intensity, frequency contrast, temporal contrast, orientation, and pitch.* A gist vector is then calculated from these features and Principal Component Analysis is performed to remove redundancy in the information contained in the gist features. The authors conclude that the proposed algorithm with gist-like features perform roughly the same as humans on stress labeling test. The referred work presents an experiment in prominence detection in speech, but the authors argue that the method can be seen as a general biologically inspired model for audition.

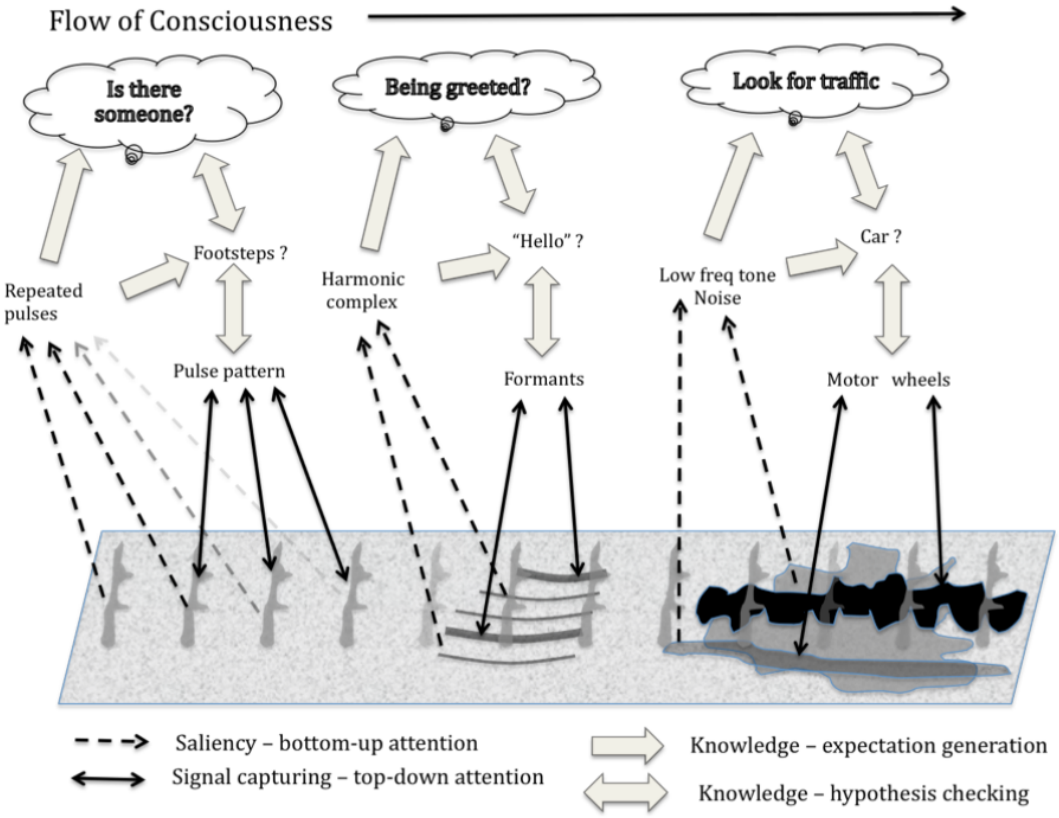

Figure 2.12: *The interplay between hearing and listening in a situation where footsteps and greetings can be heard. From (Andringa 2010)*

2.3.4 Do humans recognize auditory objects?

In the preceding sections of this chapter a theoretic framework was constituted that can account for the human ability to interpret soundscapes in complex, everyday situations. An important question remains: what are the 'units of perception' that humans recognize in the stream of information that enters the auditory cortex and how are these objects formed? In other words: what are the building blocks of our conscious auditory experience? What is the basis upon which attention selects the relevant parts of the auditory input: frequency regions, spatial locations or perceptual objects formed in early stages of auditory recognition? This section will argue that the notion of 'auditory objects' can explain auditory stimulus selection. In this view perception, attention and object formation are intrinsically linked.

Four principles for auditory object perception

The debate over the existence of auditory objects has not been resolved yet; there currently is no general theory for the perceptual building blocks of our conscious experience of sound. For vision however these principles have become accepted to some extent. (Griffiths and Warren 2004) proposes four general principles for perceptual object analysis in humans and

applies these principles to auditory perception. This review provides a good starting point of the current discussion. Below four important principles from Griffiths and Warren (2004) are articulated:

1. Object perception is the analysis of information that corresponds to identifiable objects in the 'real world'.
2. Object analysis involves the *separation* of relevant and irrelevant information: incoming sensory information that can be ascribed to objects in the sensory world is separated from information that can be regarded as 'background information' or 'noise'.
3. Object analysis involves *abstraction* of information that describes the object, so that it can be recognized. The phenomenon of object constancy in vision is an example of this: we are able to recognize an object across transformations in size, viewing angle, perspective etc.
4. Object analysis involves *generalization* between scenes: when perceiving a speaker, the voice that is heard is perceived as the same object as the visual counterpart of the percept, the moving lips face of the speaker.

Griffiths and colleges argue that the first three of these general principles are essential for explaining conscious experience. The fourth principle, cross-modal correspondence between for example the visual and auditory recognition process in looking at and listening to a speaker, is debatable.

Low-level attention or early object creation?

(Alain and Arnott 2000) in seeking an explanation that links selective attention to auditory objects, divides the existing theories in two main views on attention and object recognition:

Theory 1: Selective attention functions upon low-level features of the stimulus, emphasizing relevant frequency regions and binding together features that are likely to be related. This integration not only acts in the frequency plane; integration over time is needed for sound. These 'meaningful units' are what selective attention works upon, selection the most salient object for conscious processing.

Theory 2: Object formation is achieved without attention or awareness: a process of preliminary analysis constitutes auditory objects and further examination of each object (one at a time) is possible when selective attention processes enable further conscious analysis. In this view the experiment that Cherry (the dichotic listening task, see section 2.3.2) carried out may be explained as follows: information from both streams is integrated to one object that stands out, but conscious processing of the content of the stream can only be performed on the attended object.

Alain and colleagues then proceed to describe an experiment that seeks to capture the organization of auditory perception through event related potential measurements (ERP); the results promote the second theory, that of early establishment of auditory objects in cortical areas.

Reverse hierarchy theory

Another study, motivated by insights from neuroanatomy, promotes the early creation of auditory objects. Nelken and Ahissar (2006) applies the *reverse hierarchy theory* to audition; this theory argues that initial (conscious) perception is signal-driven, holistic and coarse. The holistic representation in the primary auditory cortex serves the need for perceiving the 'gist' of an auditory scene as discussed in section 2.3.3. When more than the 'gist' is needed, further processing takes place along a top-down path, backwards in the direction of the signal, hence the name of the theory. In this view, initially coarse objects are formed in the auditory cortex; Nelken hypothesizes that properties of these objects are calculated by higher-order cortical areas.

Despite the fact that measuring the neural response for auditory cues is more difficult than visual perception and attention (Scott 2005), this area is gaining attention. As an example we refer to (Kumar et al. 2007) that hypothesizes a model that can explain the interplay between Heschls gyrus, the planum temporale, and superior temporal sulcus. 16 models were tested against experimental fMRI data.

Shinn-Cunningham (2008) defines an auditory object as *'... a perceptual entity that, correctly or not, is perceived as coming from one physical source'*. In this article it is argued that the same attentional mechanism underlies the different attention phenomena across perceptual modalities. A key question in audition is how humans are able to interpret auditory streams from multiple sound sources at a time. Shinn-Cunningham (2008) argues that it is probable that the human auditory system attends only one source at a time and divides attention between different streams that require attention. The main problem with this theory is that it leaves the brain with incomplete stimuli: shifting attention in the auditory domain takes about 100-200 ms. The human brain pays a price for observing multiple sonic sources at a time (in the view of Shinn-Cunningham: attend to the sources serially) because from each source only the attended part of the source signal can be processed. However, the information that is missed belonging to the temporarily unattended stream can be filled in by smart guessing; an example from speech perception is phonemic restoration where missing bits of a speech stream are filled in to restore the data to form phonemes. In the case of speech, the context (i.e. surrounding the sentence) of the acoustic information is strong enough to guide recognition to a realistic hypothesis even when large parts of an acoustic stream are missing. However, perfect restoration is impossible: working memory degrades in time and therefore attending of two streams at a time is not as accurate as attending only one stream (Shinn-Cunningham 2008).

2.4 Audition: Summary and conclusion

This chapter has presented a large body of scientific work centered around perception of environmental sounds. This final section concludes the discussion of literature concerning audition.
Our discussion started with describing the history and current state of art of environmental sound recognition; we argued that despite efforts in speech recognition and music transcription, there are still major difficulties that need to be overcome for universal sound recognition to emerge. One important step is to acknowledge that sound recognition systems need to function in the *real world* and need to be able to recognize target sources in complex (and possibly mixed) soundscapes. This need should be reflected in the choice of the datasets that are used to train and validate the algorithms that underly the system. In this thesis we argue for the development of systems that can function *outside* the lab under difficult circumstances and aims at collecting and annotating recordings from complex, everyday environments that can be used to train and test recognition systems realistically.
Choosing these stimuli (called 'environmental sounds' in this work) is not straightforward. Sound researchers debate the type of sounds, the amount 'cleanness' (versus noisiness) of the recordings and the way recordings should be stored and indexed.
Also role of segmentation data in databases of environmental sounds is debated among sound researchers; this project chooses an approach that allows the annotation of sounds *within* a recording. For each sound source recognized by the annotator the timing is stored and a description is then attached to this segment.

William Gaver made a major contribution to the scientific understanding of how humans hear in the world in his two 1993 articles. His ecological account of human audition provides a good starting point in explaining the relation between listeners and their (sonic) environment. An influential attempt to understand how humans organize sound into meaningful elements came from Bregman (1990). His theory of *stream selection* however fails to explain top-down and attentional influences.

Humans audition is well tailored to segment different 'streams' of sound from complex environments: early experiments already revealed the ability to selectively attend to part of the stimulus and filter out other sounds. In recent explanations the phenomenon *attention* plays a major role in the process of selecting, integrating and segregating perceptual elements from the bombardment of auditory input. Confusingly, the term *attention* still lacks a common understanding among cognitive psychologists. At least two cognitive routes of information mediated by attention can be distinguished: signal driven attention that promotes stimuli to be selected by *saliency*, and knowledge driven attention that consists of *top-down influences*. The latter form is crucial when consciously performing a demanding task, but may severely influence overall auditory perception; attentional focus on one source, area (in the frequency plane) or task can cause received but unattended stimuli to stay unnoticed for the listener.

The discussion of different forms of stimulus omission through attentional process reveals **two important implications for an annotation task**:

1. **The task assigned to the annotator can influence perception and resulting reports of sound sources,** and therefore it is crucial to carefully design and communicate the task and annotation setup.

2. **Attentional processes structure the annotator's perception;** perception is never neutral, but always the result of a complex interplay of different processing stages and the neural correlates that perform these stages. A human report on sound sources present in a recording therefore should therefore never be considered objective. When human reports in the form of annotations are used to train and test non-human classifiers, this should be reflected in the formulation of the goal of the recognition system.

This chapter also introduced the notion of *gist*, a concept that may explain how sensory and cognitive resources are used in early perception to produce a representation of the scene that is optimized for attentional processes. Gist theory predicts that this representation contains information from all levels of processing. This includes high-level (object) information, but only to a level of detail that is nescessary to allow further selection of interesting areas. The discussion in this chapter may suggest that processes producing this gist are separate from attentional selection. This is probably not the case: it is well possible that attention is not a separate process, but an emergent property of the cognitive systems producing the gist representation and thus is intimately linked with the constitution of the gist representation.

The last section of this chapter discussed the existence and nature of *auditory objects*: what are the 'units of perception' that are extracted and created from an auditory scene during conscious perception? The *reverse hierarchy* theory is proposed as an explanation of object formation and attentional processes that also embraces the notion of *gist*.
An implication when adopting this theory for listening in an annotation task is that it provides a basis for representing people's reports on sonic events as 'objects': time segments in the recording that have clear boundaries and to which a semantic description can be attached.

2.5 Related work

Before moving on to an actual implementation of an annotation tool, we will first look at related work from different fields. This project is not the first to aim at gathering annotations for real world sound recordings; there are more projects that look into building descriptions of (environmental) sound recordings. This section provides an overview of scientific work that relates to our project. We will explore two main areas:

1. First we will look at a number of projects that aim at **collecting real-world sound recordings**.

2. Secondly, we look at a selection of projects that **develop software tools for the annotation of sound recordings** in different forms.

The goal of this section is not to provide an extensive overview of all the work in both fields, but provides a glossary of related projects and shows how this project (and, more in general, the approach of the Auditory Cognition Group in sound recognition and soundscape annotation) takes a unique position in the fields of sound recognition and human audition research.

2.5.1 Related work: Databases of real world sounds

This section provides an overview of projects that collect real-world sound recordings. For each project the most important data are presented and a short description is given.

DARES

Name:	Database of Annotated Real-world Everyday Sounds: DARES
Author:	Maarten van Grootel, Auditory Cognition Group, RUG
URL:	`http://www.daresounds.org`
N' recordings contained:	120
Attribution:	Closed
Recording type:	60 seconds long WAVE fragments
Annotation type:	Start-stop, no frequency information, extensive annotation in over 800 classes.
Publication:	(Grootel et al. 2009)

DARES is a database that contains 120 recordings of 60 seconds from everyday situations, such as the kitchen of a small appartment and a bus stop. Annotations were made directly after recording the audio, allowing the annotator to be very detailed in the descriptions. An 'artificial head' was used to mimic the human anatomy to some extend.

DESRA

Name:	Database for Environmental Sound Research and Application
Author:	Brian Gygi and team
URL:	`http://www.desra.org`
N' recordings contained:	223 recordings, in development
Attribution:	Open, registered users can upload sounds and set tags.
Recording type:	What Gygi calls 'environmental sounds': see section 2.2.3
Annotation type:	Multiple tags per recording, no within-file timing information
Publication:	(Gygi and Shafiro 2010)

DESRA is a database by the group of Bryan Gygi, currently in development. It aims at providing a resource for research in recognition of environmental sounds. In (Gygi and Shafiro 2010) the design and principles that underly the database are motivated. Currently the database contains a few hundred samples.

Freesound

Name:	The Freesound project
Author:	Multiple.
URL:	`http://www.freesound.org/`
N' recordings contained:	$>$ 100.000
Attribution:	Open, subscribers can add and edit tags.
Recording type:	WAVE recordings in different bitrates and formats.
Annotation type:	Multiple tags per recording, no timing information. Some samples are also 'GeoTagged'; the recording location is known in detail.
Publication:	None.

Freesound is a very large and open collection of audio recordings. Any internet user is free to contribute to the collection and is allowed to add and remove tags for other recordings. For every recording a set of labels is provided that describe the recording and allows for searching and grouping. Manipulated sounds are allowed in the database. Open attribution makes the labeling very diverse; samples with the same tag may contain very different types of audio recordings.
The websites `SoundSnap.com` and `SoundIdeas.com` offer similar services, but are not free, unlike Freesound.

Carnegie Mellon Auditory Lab collection

Name:	Sound Events Database
URL:	`http://stims.cnbc.cmu.edu/AuditoryLab/data-files.html`
PAGINA Author:	Laurie M. Heller, Auditory Lab of Carnegie Mellon University
Attribution:	Closed
N' recordings contained:	?
Recording type:	'clean' recordings of typical sounds in the categories 'impact, deformation, rolling, air, liquid events'. WAVE
Annotation type:	Very short description in text
Publication:	(Heller and Skerritt 2010)

This database provides a set of sound event recordings made for research purposes. Details on the process of producing these sounds are provided, including videos that show how the sounds were produced.. Recordings were made under controlled conditions, limiting background noise and influences of room accoustics.

2.5.2 Related work: Other tools for multimedia annotation

This section discusses related work on the annotation of sound recordings. Four examples are discussed: ELAN, IBM multimodal annotation tool, Marsyas/Cantillion and ProjectPad.

ELAN

Name:	EUDICO Linguistic Annotator
URL:	`http://www.lat-mpi.eu/tools/elan/`
Author:	Max Planck Institute for Psycholinguistics
Target media:	Multimedia annotation of recordings that contain language
Annotation type:	Segments in time, hiearchical annotation possible.
Output format:	XML, and exports to Shoebox/Toolbox, CHAT, Praat and text files
Platform:	Java (Windows, Linux, MacOS)
Publication:	(Wittenburg et al. 2006)

ELAN is an annotation tool that aims at supporting the annotation of multimedia recordings that contain language of any kind, including sign language. It allows the annotator to enter different levels of description, including syllable-level annotations. A screenshot can be found in figure 2.13. ELAN is widely used in language research.

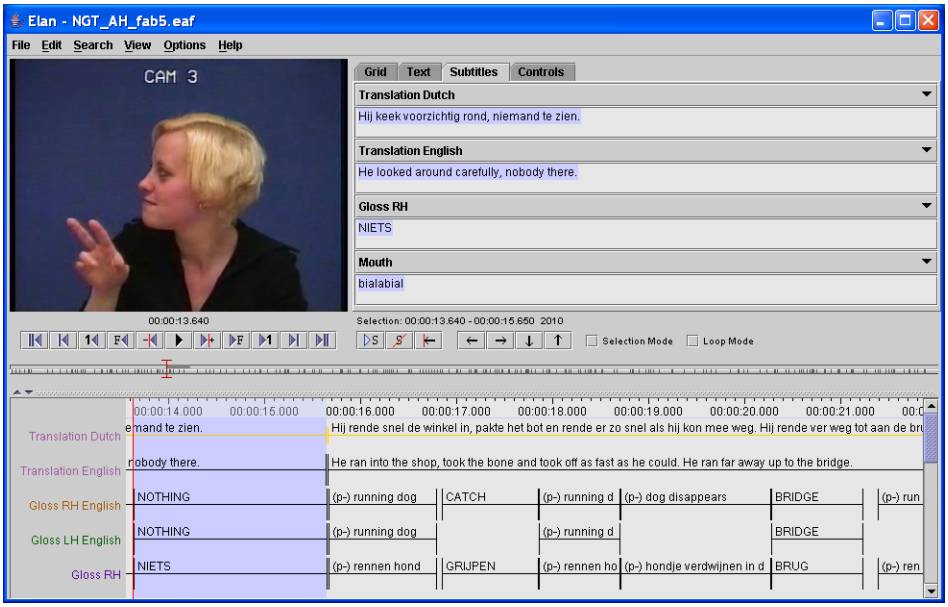

Figure 2.13: *Screenshot from the ELAN annotation software, built to annotate multimedia recordings containing sign language.*

IBM Annotation Tool

Name:	IBM Multimodal Annotation Tool
URL:	`http://www.alphaworks.ibm.com/tech/multimodalannotation`
Author:	IBM, Bill Adams
Target media:	MPEG-7 data
Annotation type:	Semantic annotations
Output format:	XML
Platform:	Windows
Publication:	(Lin et al. 2003)

'A tool that assists in annotating MPEG files with MPEG-7 meta data; both video and audio annotations can be created.' This IBM project has been suspended since 2002. VideoAnnEx [3] is the video annotation counterpart of this project that is still available online. This tool allows users to annotate both scenes ('shots') and parts of those scenes, and it implements some assistence guided by MPEG-7 descriptors.

[3] http://www.research.ibm.com/VideoAnnEx/

Marsyas, Cantillion

Name:	Marsyas: Music Analysis, Retrieval and Synthesis for Audio Signals
URL:	`http://marsyas.info/`
Author:	George Tzanetakis
Target media:	Audio
Annotation type:	-
Output format:	Depends on application
Platform:	Windows, Unix (Linux and MacOS)
Publication:	(Tzanetakis and Cook 2000c)

This is not a complete annotation application, but a framework that allows researchers to build audio analysis and visualization systems. One example of how this project can be used to support an annotation paradigm is Cantillion (`http://cantillion.sness.net/`), a web-based application that allows researchers to annotate recordings of *chants*, religious songs. Annotation is performed on a time line that also presents the pitch of the chant.

ProjectPad Audio Tools

Name:	ProjectPad Audio Tools
URL:	`http://dewey.at.northwestern.edu/ppad2/07timeline.html`
Author:	Northwestern University, Evanston (Illinois)
Target media:	Audio (mp3) and video (Flash)
Annotation type:	Timeline
Output format:	XML format
Platform:	Java
Publication:	-

The Video and Audio Tools lets you attach comments to time segments of FlashFLV video and MP3 audio streams. The tools can be used by instructors and / or student teams to critique student-produced video and audio or to provide a way for students to analyze scientific, historic, or artistic recordings. The application allows the user to store notes (annotations) on different axes of a timeline. The underlying ProjectPad application is a web-based multi-user environment that allows multiple 'annotators' to work together on the description of a media source.

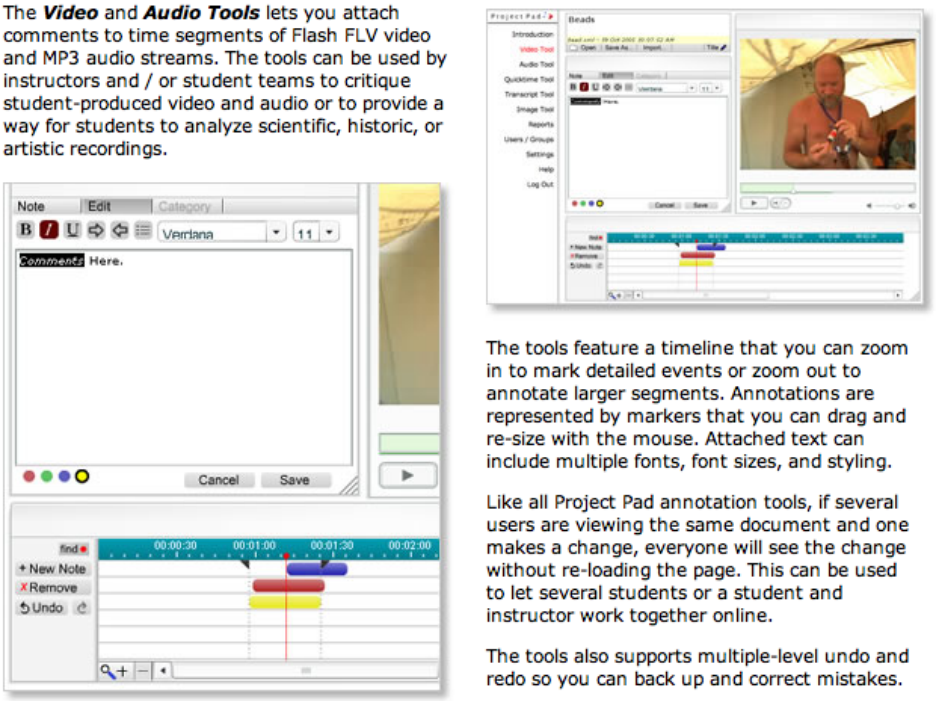

Figure 2.14: *Screenshots of Projectpad, a Java sound annotation application. The application is also capable of playing Flash videos for annotation.*

Chapter 3

Implementing a tool for annotating sound files

In the introductory chapter of this thesis we presented the research objectives for this project. A major objective is to develop a novel implementation of a software tool for soundscape annotation. This chapter discusses implementation choices and details for this tool.

3.1 Design choices

This section presents the design choices that were made in the development of the annotation tool. We will discuss why a cochleogram was used and how it was preprocessed. Section 3.3 presents the technical requirements and actual implementation of the different windows and dialogs.

3.1.1 Cochleogram representation

A cochleogram image (Andringa 2002) was added to the interface to allow the user to visually inspect the audio stream, identify (un)interesting segments and allow rudimentary identification for a trained eye. In principle a cochleogram image is derived from a model of the cochlea (see the Background chapter), that resembles a Fourier transform to some extent. To this representation two models are applied that track the parts of the signal in the temporal-spectral plane that will be perceived as the *foreground* of the soundscape, and the part that will be regarded as the *background*. These to images are superimposed and result in a representation that shows clear separation between the foreground and background part of the soundscape.

3.2 Previous work: MATLAB version of the tool

An earlier version of the annotation tool was implemented in Matlab. This tool was also used in the experiment described in (Krijnders and Andringa 2009a). While Matlab is suited

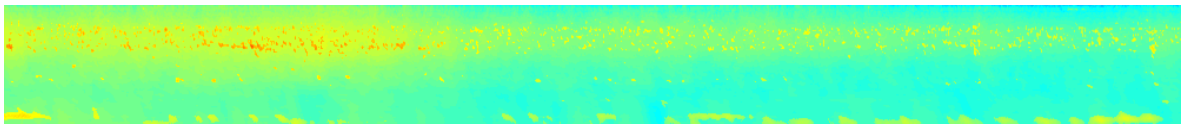

Figure 3.1: *Fragment of a cochleogram as it was presented in the annotation tool to provide visual feedback for the annotator. The colors were slightly edited to improve readability in print.*

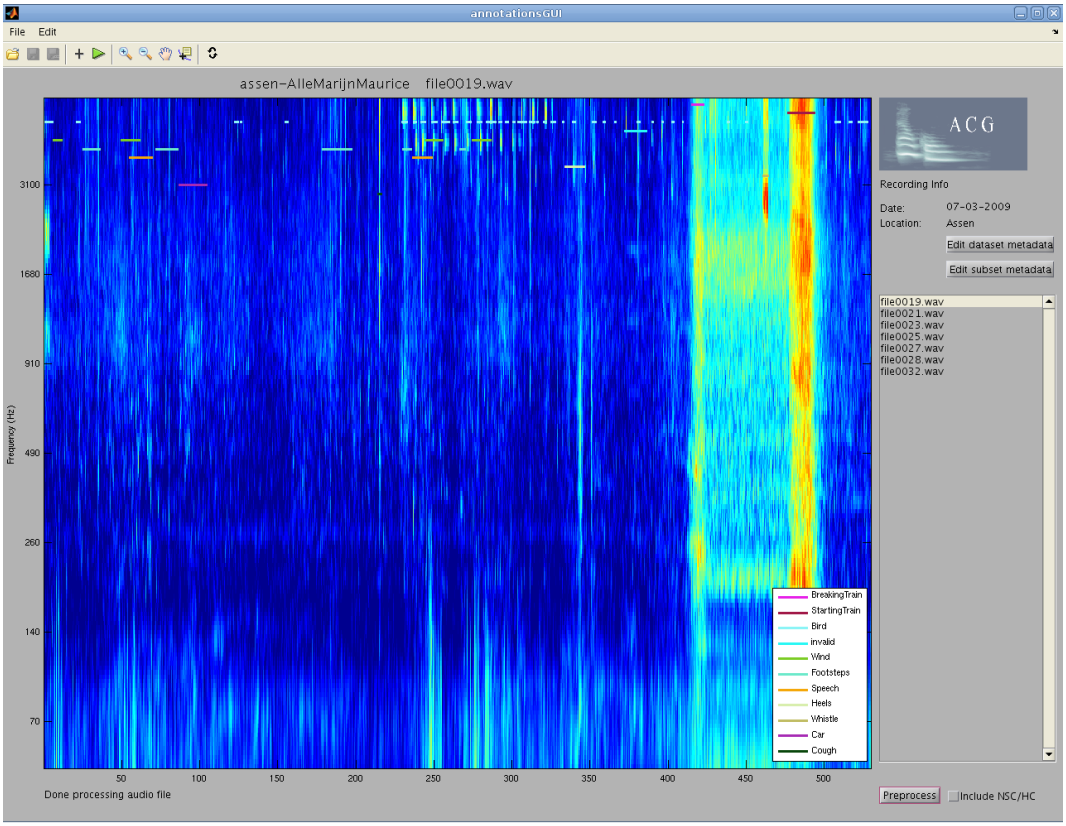

Figure 3.2: *Screenshot of the Matlab version of the annotation tool. The annotations are shown as colored bars on top of the cochleogram. Audio playback functionality is provided by the buttons in the toolbar.*

well for processing and analyzing audio data and annotations, its capabilities of running a graphical user interface are limited. When implementing a Matlab annotation tool it showed that the graphical interface becomes slow when drawing a large number of annotations of the screen.
To overcome these limitations, it was chosen to develop a new implementation of the tool in a scripting or programming language that does not have these limits. This chapter describes this process, in the next chapter describes an experiment that tests the annotation tool.

3.3 Development of soundscape annotation tool in Python

A new, stand-alone and lightweight annotation tool was developed in Python. This section will describe the design considerations and implementation details of this new tool.

```
<event>
  <start>46.7719891071</start>
  omni-directional1786</stop>
  <type>Traffic</type>
  <recordingUID/>
  <annotator>ray</annotator>
  <location>
    <x>NaN</x>
    <y>NaN</y>
    <z>NaN</z>
  </location>
</event>
```

Figure 3.3: *Example (fragment) of an XML file. An Event entry respectively holds the following information: the boundaries for the time segment that is annotated (in seconds), the type of annotation (a textual description), a unique ID for this event, the name of the annotator, and unused variables to store the location of the event in the case of multi-channel recordings.*

3.3.1 Annotations output format

Annotations were saved for later retrieval in an XML document that follows a pre-defined format [1]. XML was chosen because this file format human-readable, which allows for easy inspection and debugging. In the dataset file fields are reserved for common descriptors such as the recording location, recording device (with its characteristics) and the time and date of the recording. As an example this format a fragment is given below that represents one annotation of an event. The 'location' fields are currently not used but might be used in the future to store the coordinates of the sounds source, derived from omnidirectional microphones.

The XML file also holds references to the corresponding audio files; in a multi-channel recording the signal from each channel should be contained in a separate WAVE file. An XML file can contain multiple datasets, but all datasets in that file share the same list of classes.

3.3.2 Ontology

For the experiment described in the next chapter the goal was to establish how annotators vary in both their choice of annotations and class labels. Therefore the current version of tool allows the annotator to choose class labels freely; no pre-selected ontology or class list was implemented. Future versions of the tool may include this option, see section 6.2.6 for a discussion of the options for future research.

[1] Available from `http://www.daresounds.org`.

3.3.3 Annotated sound datasets in use at ACG

The annotation tool that is developed should be able to load all types or real-world sound datasets that are currently in use at the ACG group. These datasets are listed below. All these datasets contain environmental sounds (see chapter 2 for a definition).

Amstel dataset from Cassandra (Zajdel et al. 2007) project: 1 hours of sound recordings from an experiment in aggression detection. These recordings were taken on a subway station, Amsterdam Amstel, and feature professional actors who simulate normal and aggressive behavior. These recordings were extensively annotated by a single annotator.

The Database of Annotated Real-world Everyday Sounds , or Dares-G1: described in (Grootel et al. 2009): a database containing 2 hours of real-life situations. Both indoor and outdoor recordings are included.

Assen 2009 recordings : five recordings of three minutes taken by students at various places in the town of Assen. Annotations were made by various students, multiple annotations exist for the same recording.

Assen 2010 recordings : two recordings taken in a real-life street scene in the small town of Assen. One recording was taken from a busy intersection with construction nearby, the other recording was taken next to a more quiet road. Both these recordings were taken as part of an experiment described in (Krijnders 2010). Annotations are available.

3.3.4 Technical and usability requirements

To obtain an annotation tool that fits our needs, the following **usability requirements** need to be fulfilled :

1. Provide freedom for the user to develop her own strategy.
2. Provide good visual feedback: include a visual representation of the sound.
3. Provide a user interface that adheres to people's knowledge on software: 'make that button do what users expect it to do'.

There are also **technical requirements** that the tool needs to meet:

1. Valid XML output, complying the standard developed by the Auditory Cognition group. A .xsd XML schema file is available.
2. Capable of reading WAVE sound data files.
3. Cross-platform support.
4. Error reporting to track and prevent errors.
5. Facilities to log the clicks and keystrokes of the user, for later inspection.

The next chapter describes an experiment in which the tool was tested.

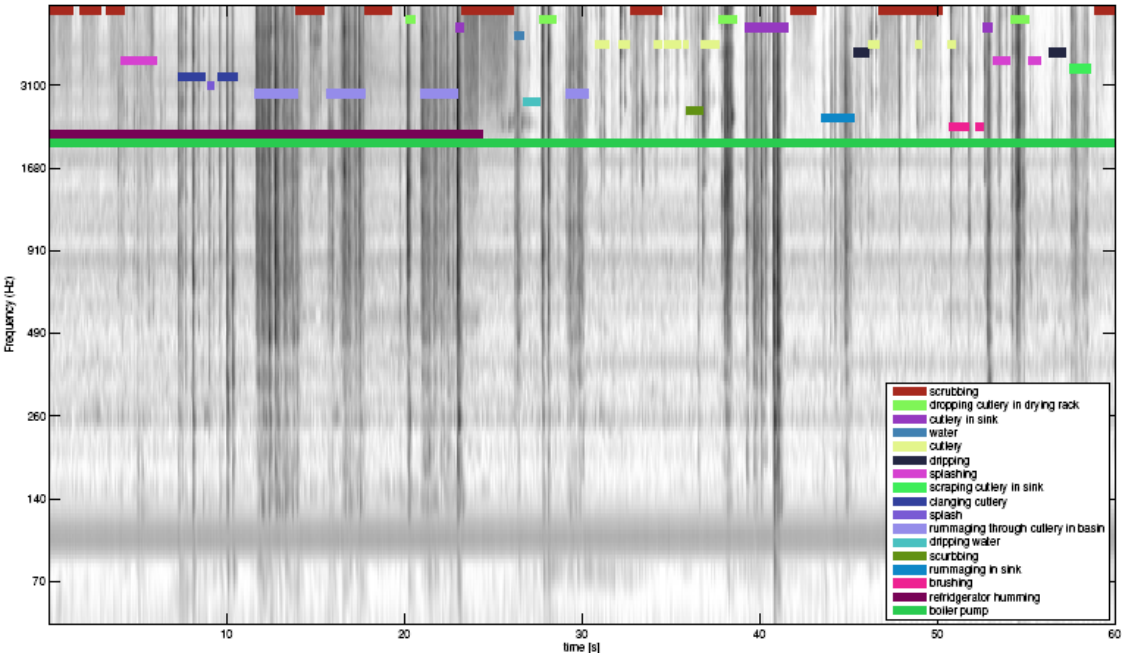

Figure 3.4: *Annotations plotted on top of a cochleogram, as the Matlab annotation tool presents them, see the screenshot on the previous page. The y-position of the annotation bar does not relate to the frequency plan, and the annotation bars may clutter the cochleogram image. In the current implementation of the annotation tool the annotations and cochleogram are plotted on separate panels.*

3.3.5 Implementation details

The application was developed in Python, a language suited for rapid prototyping. The graphical user interface uses TkInter, an API that allows the programmer to easily create Tcl/Tk windows and implement widgets. For audio playback the PortAudio API was used, interfaced to python by PyAudio. The program uses the standard Python DOM module to read, write and edit XML files. This last module is easy to use for the programmer, but a major drawback is that the XML files resulting from it are not optimized to be read by humans (i.e. not formatted).

The application was implemented and tested on an Ubuntu Linux system, but has also shown to function on Apple Mac computers and under the Windows operating system.

3.3.6 Annotation application: User interface

This subsection describes the most important aspects of the annotation tool.

The main application window

The interface consists of four major parts:

1. To the left: a **legend** indicating the class list and the colors for the corresponding annotations. The number of annotations for that class is presented in the colored box.
2. The **main annotation window**: this presents the actual annotations as colored rectangles. The portion of the signal that is shown corresponds to the portion presented in the cochleogram window. The red line indicates the current position of the audio.
3. **Cochleogram window**: always presents the full frequency range of the cochleogram, zoom level of the x-axis depends on the zoom actions of the annotator.
4. A panel containing **control buttons and information displays**. From the top: a window presenting a full overview of the signal that indicates which portion of the signal is zoomed into. Below that is a panel displaying the registered name of the annotator. Next is the audio playback panel (see 3.7) that allows the annotator to start, pause and stop the audio playback. The 'mode' panel lets the user choose between either the 'zoom' mode and 'annotate' mode, to choose between inspecting the cochleogram or adding annotations,respectively. The Zoom buttons below that allow to zoom to a chosen selection of the signal of zoom out to a full view on the signal and annotations. The 'Select a dataset' panel lists all the Dataset items that are present in the currently loaded .xml file. The bottom button provides a shortcut to store the annotations into an .xml file.

Entering an annotation

The annotator can ad an annotation by first selecting the 'annotate'-mode, then create a selection in the cochleogram window by dragging the mouse over it. As soon as the mouse button is released the dialog in figure 3.6 pops up. The annotator is presented with a cochleogram representation of the selected signal region and is asked to select the appropriate class from the list, or add a new class to the list.

Audio playback buttons

There are three buttons that allow the annotator to control the audio playback; the complete panel can be seen in figure 3.7. The application allows the user to start playback in two ways: either by continuing the playback at the current position of the cursor (i.e. the red line in the annotations window) or starting the playback at the beginning of the part of the signal that is currently selected in the zoomed window. The user cannot move the cursor directly. While this is merely an implementation issue, it enforces the annotator to move the audio visualization to the part that he/she currently listens to which may lead to better annotations.

3.3.7 Experimental software

The software that is used for the experiments described in the remainder of this chapter is experimental and not bug-free: however, before starting the experiment it was assured that no

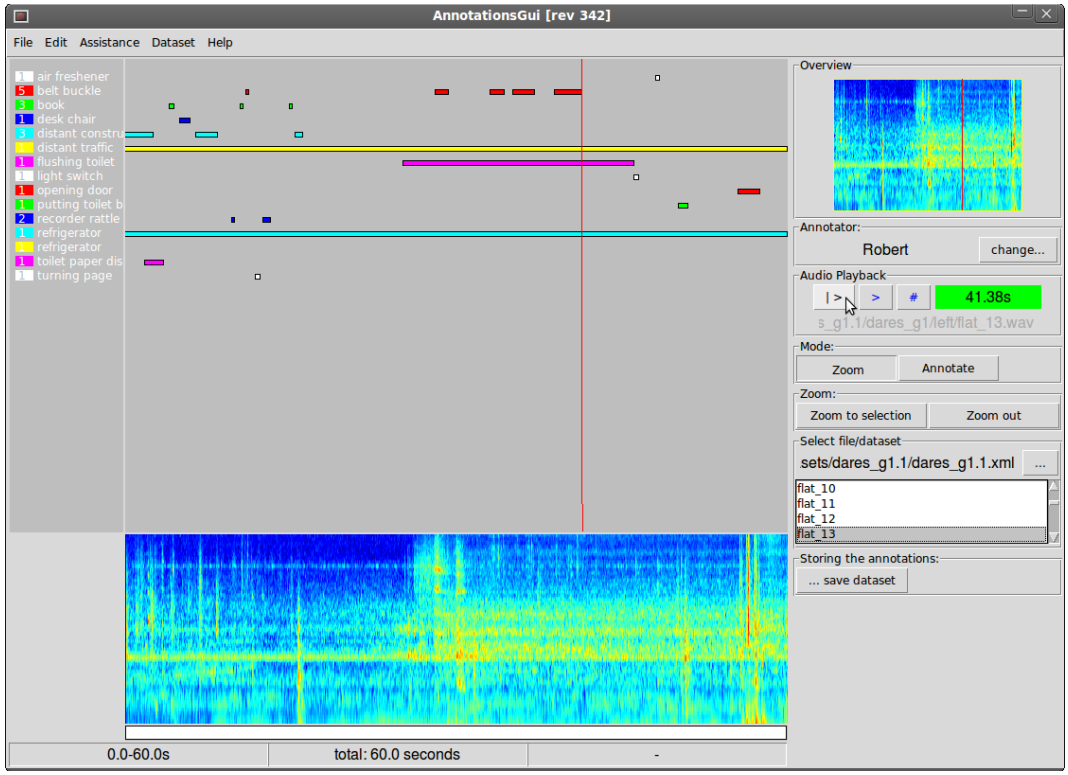

Figure 3.5: *Overview screenshot of the main window of the annotation tool.*

major errors would disturb the trials. In the near future an updated version of the software will be available for download from the DARES website, `http://www.daresounds.org`.

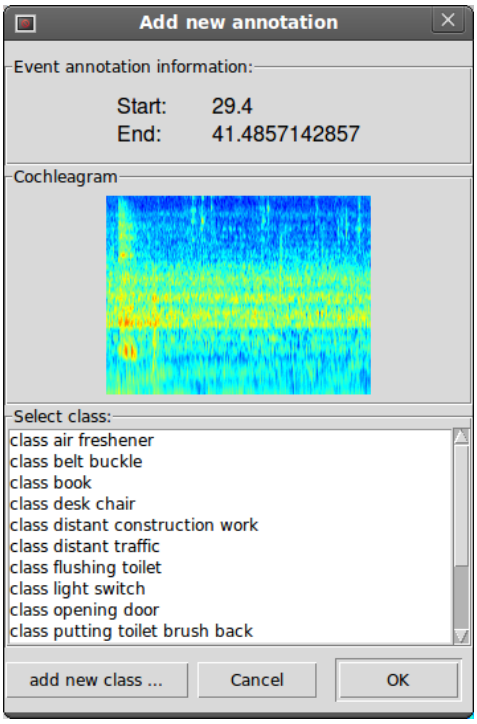

Figure 3.6: *Selection screen: the annotator is asked to select a class description from the list.*

Figure 3.7: *Closeup of the buttons used to control the audio playback. This panel features three buttons and two displays to present the position in the audio file, and the filename of the currently loaded WAVE file.*

Chapter 4

Experiment - Method

In this chapter we describe how the tool for annotating sound files was tested in an annotation experiment. The purpose of this experiment is threefold:

1. To consolidate the tool and to confirm that no critical errors show up during an annotation session.
2. To analyze how subjects perform their task; what strategies do they employ? How do the annotations differ between subjects?
3. By directing the subject to a specific strategy it was aimed to direct subject behavior into two types of perceiving and analyzing the soundscape presented to them, reflecting the difference between hearing and listening.

4.1 Method

This section describes an experiment that tests the annotation tool in general, and the specifically the behavior of annotators under time restrictions.

4.1.1 Dataset: Soundscape recording

The soundscape recording used in this experiment was recorded in Assen, a 65.000 inhabitants town in the north of the Netherlands, in March 2010 on a cold, cloudy afternoon. The recording setup consisted of a single-channel omni-directional microphone placed on a tripod connected to a digital audio recorder. The data was recorded at a sampling rate of 48 kHz and 24 bits per second. The recording setup was located in the front yard of a three story villa, a few meters away from a lively road where pedestrians and all kinds of traffic can be expected to pass. There was a hump in the road located near the recording place that is intended to slow cars down. Figure 4.1 shows photographs of the recording location.
While recording a on-line annotation experiment (performed in silence) took place which is described in chapter (Krijnders 2010); for this experiment we assume that this did not influence the recording.
The recording was cut in two equal portions, each to serve as a stimulus.

4.1.2 Subjects

Subjects were recruited among (ex-)student population of the University of Groningen, aged between 18 and 26 year old. They were not trained in annotating soundscapes or otherwise

Figure 4.1: *Left: Birds eye view of the recording location for the soundscape recording used in the experiment. The microphone was located at the red dot near the center of the picture. Right: Street-level photograph of the recording location. Copyright: Bing Maps, Microsoft.*

trained to analyze audio recordings in any way. Computer experience ranged from moderate to experienced user.
Normal hearing capabilities were a prerequisite to take part in the experiment. Participants were not obliged in any way to take part in this experiment. About half of the subjects received a financial compensation of 5 euros for their participation, the other half received credits for a course on perception held at the university of Groningen.
Three participants did not have Dutch as their native language (but Chinese, Spanish and German, respectively), the instructions therefore were given to them in English. One native Dutch participant also received instructions in English.

4.1.3 Conditions

Two conditions were created:

1. In the first condition the time provided to the subjects to annotate the recording is equal to about the length of the recording; this implies that only the most prominent sound sources can be annotated and pausing the playback of audio inherently disables the subject to listen to the whole recording.
2. In the second condition the subjects the trial time in which the annotation had to take place was set to twice the length of the recording. This enables the subject to listen to interesting portions of the recording twice and to provide more detail in the annotations.

For this experiment we assume that both recordings are equal regarding their content (because both were made subsequently and no important changes took place in the environment during recording).
We hypothesize that there is a training effect for naive subjects; no training phase is applied, therefore the two conditions are applied to both recordings in two conditions. The training effect on annotations or strategies is not measured here; this would require two more conditions.

4.1.4 Instructions

After a short introduction to the task and the recording that is used, the software tool was introduced to the subjects; the instructor showed the subject how to scroll through the recording and how to add a new annotation to the annotations frame. Initially, there were no classes or annotations given. The subjects were told to listen carefully to the recording and add annotations corresponding to the sources they identified in the recording. The subjects were told that they were free to add classes; no constraints were given to the number classes (however the software was limited to displaying 23 classes in the annotation window). Furthermore, the subjects were told that he/she could start or stop the playback of the audio at any time, and that the red line in the annotations window indicates the current position of the audio playback. The functionalities for zooming and scrolling were introduced to the subjects by means of a short demonstration. The participants were then allowed to try the application before starting the actual annotation session.
No information was given on the nature of the recording, but the subject was informed that the recording was from some random outdoor location in a real world setting and no artificial editing was performed on the recording.
The participants were told in advance that they were asked to annotate two recordings, in two different conditions. These conditions are discussed in section 4.1.3.

4.2 Data

4.2.1 Annotations

The annotations created by the subjects were stored in .XML format. For further details on this format, see figure 3.3. The user-created class list was also stored with the annotations.

4.2.2 User action registration

During the annotation session information on the behavior of the subject was automatically registered in a log file for later analysis. For each behavioral event a timestamp and description was stored. The log file holds the following information:

Audio playback information: when did the user start/pause/stop the playback of the audio and at which audio data frame did the playback start or end?

Zoom actions: When did the user perform a zoom action to examine the cochleogram? The exact portion of the cochleogram that was visible is stored.

Annotation actions: When did the subject enter an annotation? Combined with the annotations that resulted from the session it can be inferred which annotation was placed at this point in time.

Errors: if there were (software) errors during the session, these are also stored in the log file. This allows the researcher to exclude runs in which a critical error might have influenced the experiment. This information is used to debug the application and is not taken into account in the current analysis.

4.2.3 Survey

After finishing the annotations for each recording, participants were asked to fill out a survey in which they are asked to report their experience with the annotation task and the tool. The results of the survey will be presented in the next chapter.

The questions were the following:

1. Please describe your strategy in carrying out the annotation task.
2. Did you carry out the task with care?
3. Did the software work well? What is good and what could be better about the software? (please name both)
4. How did you experience the task? Was it fun? Was it tiring?
5. How sure are you that the sound sources you recognized in the recording were really there?

Dutch participants were asked to fill out a translated version of the survey, see **??** for the Dutch survey.

The next chapter presents the results of this experiment.

Chapter 5

Experiment - Results

The previous chapter described the setup of an annotation experiment; this chapter presents the numerical and qualitative results for the this experiment. In the present study twenty-one subjects (11 male, 10 female) were asked to annotate the soundscape recording described in the previous chapter using the annotation tool that was developed in this project.

The results of the experiment described here fall into three categories:

Annotation data: The sets of annotations composed by each annotator. These data are stored in .xml annotation files, but can also be obtained from the .log registration files.

User data: The sets of user data (clicks, actions, information from the audio module) that was generated during the annotation session of each participant. Section 5.3 discusses these results. These data are stored in .log registration files.

Survey answers: The results of the survey that was held among the annotators. These data are presented in section 6.1.

Before presenting the results listed above, the next section will start with describing the pre-processing of the data.

5.1 Data processing

5.1.1 Exclusion of trials

21 persons participated in the experiment. From these trials, the data of two subjects was removed after inspection. Trials were removed if:

1. The subject clearly misunderstood the instructions,
2. If the descriptions (labels) of the annotations contain descriptions corresponding to multiple classes,
3. If the subject completed the annotations for less than half of the recording length,
4. If critical disturbance of the experiment occurred.

Data from one male subject was removed because criterion 2 was applicable, another female violated criterion 1 and the data corresponding to here trial were therefore also left out of further analysis.

5.1.2 Conditions

The stimuli and conditions for this experiment are motivated in section 4.1.3. The recording was cut into two equal portions $part1, part2$, each subject only annotates each part once (to prevent training and memory effects) and is given 10 minutes to annotate one recording and 20 minutes for the other. For completion the combinations of conditions and recordings are listed below:

Condition A: The participant annotates part 1 for 10 minutes, then annotates part 2 for 20 minutes.

Condition B: The participant annotates part 1 for 20 minutes, then annotates part 2 for 10 minutes.

Nine subjects were exposed to setup A, the other 10 to setup B.

5.2 Results: Annotations

The annotation data stored in the .xml-file can be analyzed in many different ways. The next subsections discuss both the qualitative and quantitative results of the experiment.

5.2.1 Quantitative analysis

The most straightforward way to measure subject performance is to count the number of annotations the subject added during an annotation session. This data is plotted in figure 5.1, on the next page.

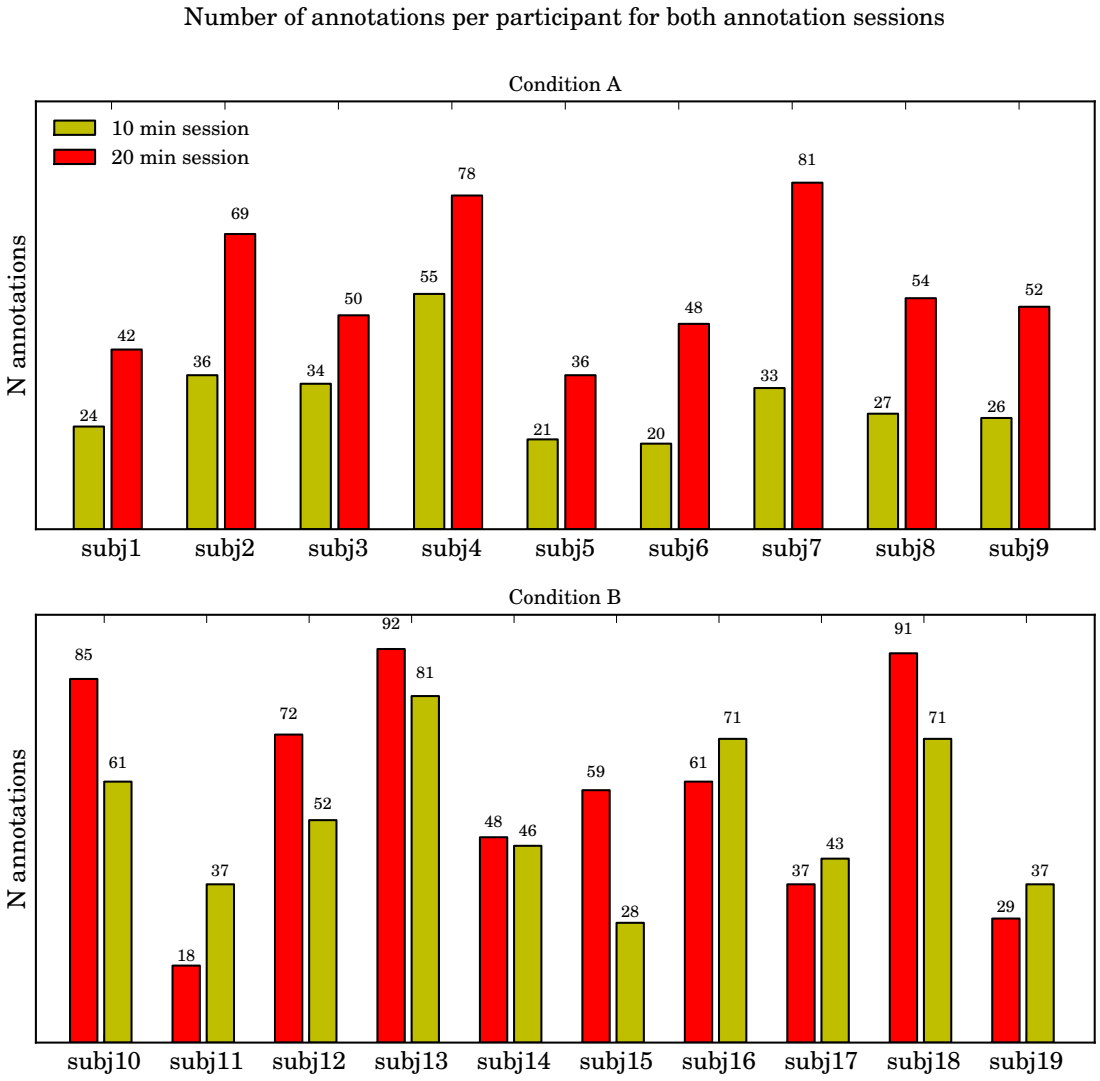

Figure 5.1: *Number of annotations per participant, for 21 participants. The numbers in the participant's names do not reflect the order in which the corresponding experiment was carried out. For each subject the lines correspond to the annotation sessions performed on recording* $\{part1, part2\}$*, respectively. The numbers on top of the bars indicate the height of that bar.*

Table 5.1 summarizes this data for all participants by presenting the mean annotation count and standard deviation for each condition-recording pair, see figure 5.2 for a plot of these data. The number of eventually deleted annotations is also included in the plot; these data were obtained from the .log file discussed below. Please note that the length of the recording sessions alternates between conditions.

5.2.2 Choice of classes

In total, subjects added 189 classes to describe the contents of the two recordings. Upper- and lowercase instances were counted as one class description. The class 'auto' (car) was the most common among the descriptions attached to annotations.

Number of classes

Annotators differ in the number of (original) classes they assign to describe the sound sources they report in the recording.

To make annotations comparable between subjects, a list of 'common' classes was constituted:

```
{'aircraft', 'bang', 'bicycle', 'birds', 'bump', 'bus', 'car', 'dog',
 'footsteps', 'horn', 'motor', 'music', 'people','rustle', 'scooter',
 'traffic', 'train','truck'}
```

The participants' classes were then mapped to a class from the 'common' list, by picking the closest matching class by hand. Some classes fitted none of the common categories and were dropped. The mapping is reported in the appendix.

5.2.3 Annotation frequencies per 'common' class for recording Part 1

To study the number of annotations per category, the annotations for each category were collected from the .xml annotations files for all annotators, grouped per condition. These data were analyzed separately for each annotation session (and therewith for the corresponding recording). Table 5.2 presents the mean number of annotations for each class, together with the standard deviation, for both conditions. Figure 5.3 presents these numbers graphically.

5.2.4 Annotation frequencies per 'common' class for recording Part 2

For the second part of the recording we present the same data as on the previous page, now for the second annotation session.

5.2.5 Visualizing annotations

A script was used to plot a timeline visualization of the .xml annotation files. The script was written in Python and invokes Matplotlib. These plots allow visual inspection and comparison of the annotations. See figure 5.5 for an example.

	Part 1		Part 2	
	20 min	10 min	20 min	10 min
Persistent annotations	59.20	30.67	56.67	52.60
Removed annotations	1.40	0.56	1.44	0.20

Table 5.1: *Mean annotation counts for each condition (20 minutes, 10 minutes) and recording (Part 1, Part 2) pair.*

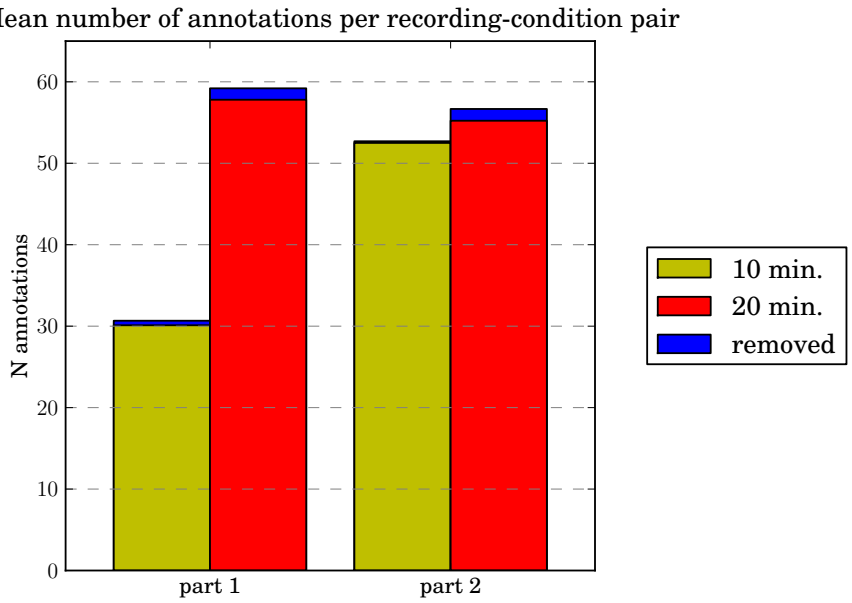

Figure 5.2: *Graphical presentation of the data in table 5.1. The blue bars represent the mean number of annotations that were removed during the trail. Note that the first and the fourth bar correspond to condition A, the second and third bar to condition B. The number of deleted annotations was extracted from the log file, the other data is calculated from the annotation files.*

Class	10 min. annotation Mean	(std)	20 min. annotation Mean	(std)
aircraft	0.00	0.00	0.40	1.20
bang	0.44	0.50	0.80	0.60
bicycle	1.00	1.05	2.80	3.46
birds	3.89	3.73	14.50	12.20
bump	1.00	2.21	4.80	6.72
bus	0.11	0.31	0.00	0.00
car	19.44	5.17	21.10	10.62
dog	0.00	0.00	0.10	0.30
footsteps	0.00	0.00	1.20	1.83
horn	0.00	0.00	0.00	0.00
motor	0.11	0.31	0.30	0.64
music	0.11	0.31	0.10	0.30
people	2.11	0.74	4.80	2.14
rustle	0.00	0.00	0.20	0.40
scooter	0.00	0.00	0.40	0.92
traffic	0.00	0.00	1.20	2.75
train	0.00	0.00	0.20	0.40
truck	1.89	1.91	3.90	8.17

Table 5.2: *Mean and standard deviation of general class frequencies for two conditions (10 minutes annotation and 20 minutes annotation).*

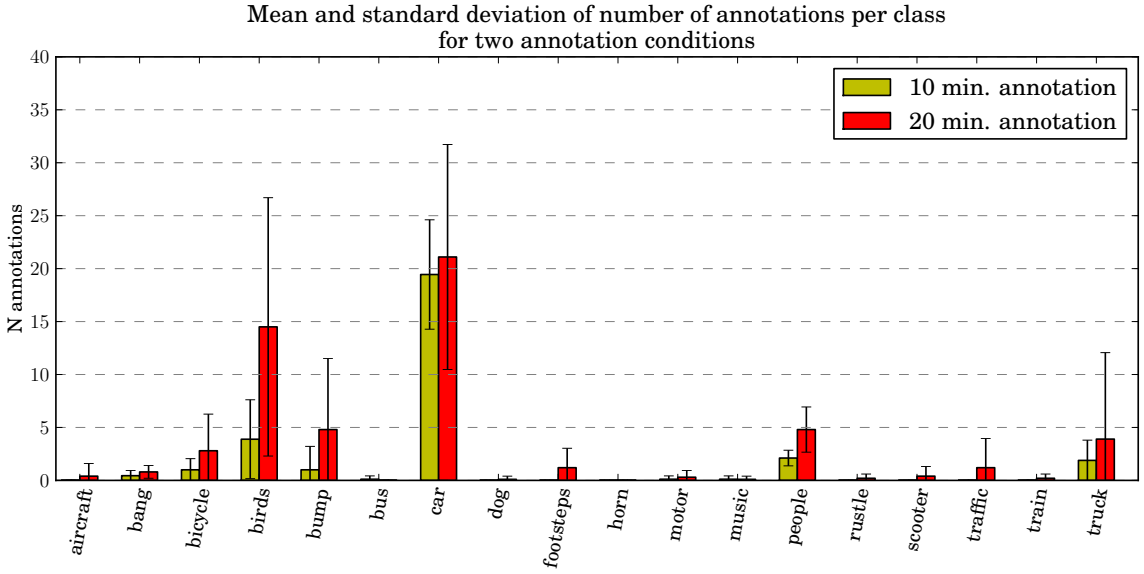

Figure 5.3: *Plot for recording 1: Mean and standard deviations for frequencies of occurance of grouped classes. These data were extracted from the XML-files containing the annotations for each subject.*

Class	10 min. annotation Mean	(std)	20 min. annotation Mean	(std)
aircraft	1.60	1.02	2.22	1.75
bang	0.80	0.98	1.22	0.63
bicycle	3.30	3.82	1.89	1.59
birds	11.90	8.87	6.44	6.77
bump	4.50	8.33	0.00	0.00
bus	0.00	0.00	0.11	0.31
car	21.60	4.65	28.22	6.89
dog	0.00	0.00	0.00	0.00
footsteps	1.00	0.77	1.67	1.25
horn	0.20	0.40	1.00	0.82
motor	0.90	1.45	0.44	1.26
music	0.00	0.00	0.00	0.00
people	3.00	1.67	5.22	2.25
rustle	0.70	0.90	0.33	0.47
scooter	1.60	1.11	2.11	0.99
traffic	1.00	2.10	0.22	0.42
train	0.00	0.00	1.33	3.77
truck	0.10	0.30	2.44	3.44

Table 5.3: *Mean and standard deviation of general class frequencies for two conditions (10 minutes annotation and 20 minutes annotation).*

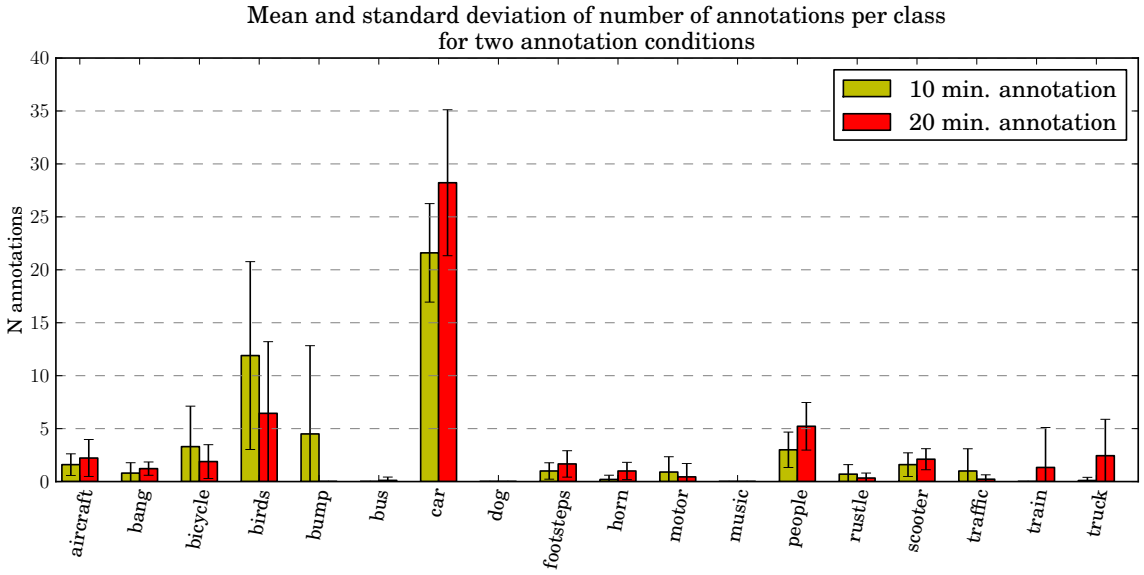

Figure 5.4: *Plot for recording 1: Mean and standard deviations for frequencies of occurrence of grouped classes. These data were extracted from the XML-files containing the annotations for each subject.*

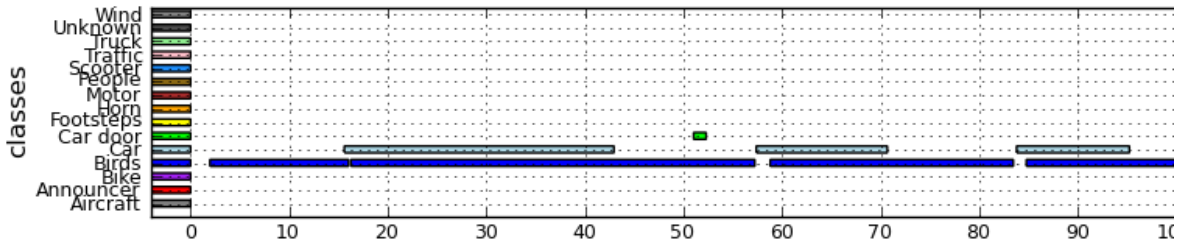

Figure 5.5: *Example fragment of semantic annotations plotted on a time line representing. The x-axis represents the elapsed time of the soundscape recording. The order in the y-axis is the order in which the subject added the classes.*

5.2.6 Combining annotations: confidence on soundscape contents

To analyze the agreement between annotators it is useful to have a way of 'adding up' the annotations of these different listeners. We propose a way of 'summing' annotations that provides an overview of the occurrence of a sound event belonging to a class for each point in time in the recording. The number of annotators making an annotation for a specific class can be seen as a measure of *confidence* for the existence of that sound event. To examine the effect of the time restriction (condition A and condition B) this measure is calculated and plotted for both conditions separately. These data were produced and visualized by applying the following steps to the annotation data:

1. Select the annotation sets of all participants that belong to the chosen recording.
2. For each participant, gather the annotations for the chosen class.
3. Constitute intervals: divide the recording in bins of 1 second length, thus creating about 610 bins.
4. Loop over bins, and for each bin check for each participant if an annotation of the chosen class is present. Store the number of annotations found for each time interval.
5. Normalize the number of annotations by dividing them by the number of annotators. This is the 'confidence measure'.
6. Repeat from step 2 for both conditions.
7. Create a smoothed dataset by running a Hanning window over the bins.
8. Plot in a figure:
   (a) the confidence measure for annotations for each condition,
   (b) the sum of the two plots from both conditions,
   (c) the smoothed curve for the summed data.

Most participants did not manage to annotate the whole recording for both conditions, therefore the plot was cut off at 480 seconds into the recording.

The next sections present the plots for each class for the two recordings.

Confidence plots for Part 1 of the recording (first annotation session)

See figures 5.6, 5.7, 5.8 and 5.9.

Confidence plots for Part 2 of the recording (second annotation session)

See figures 5.10, 5.11, 5.12, 5.13.

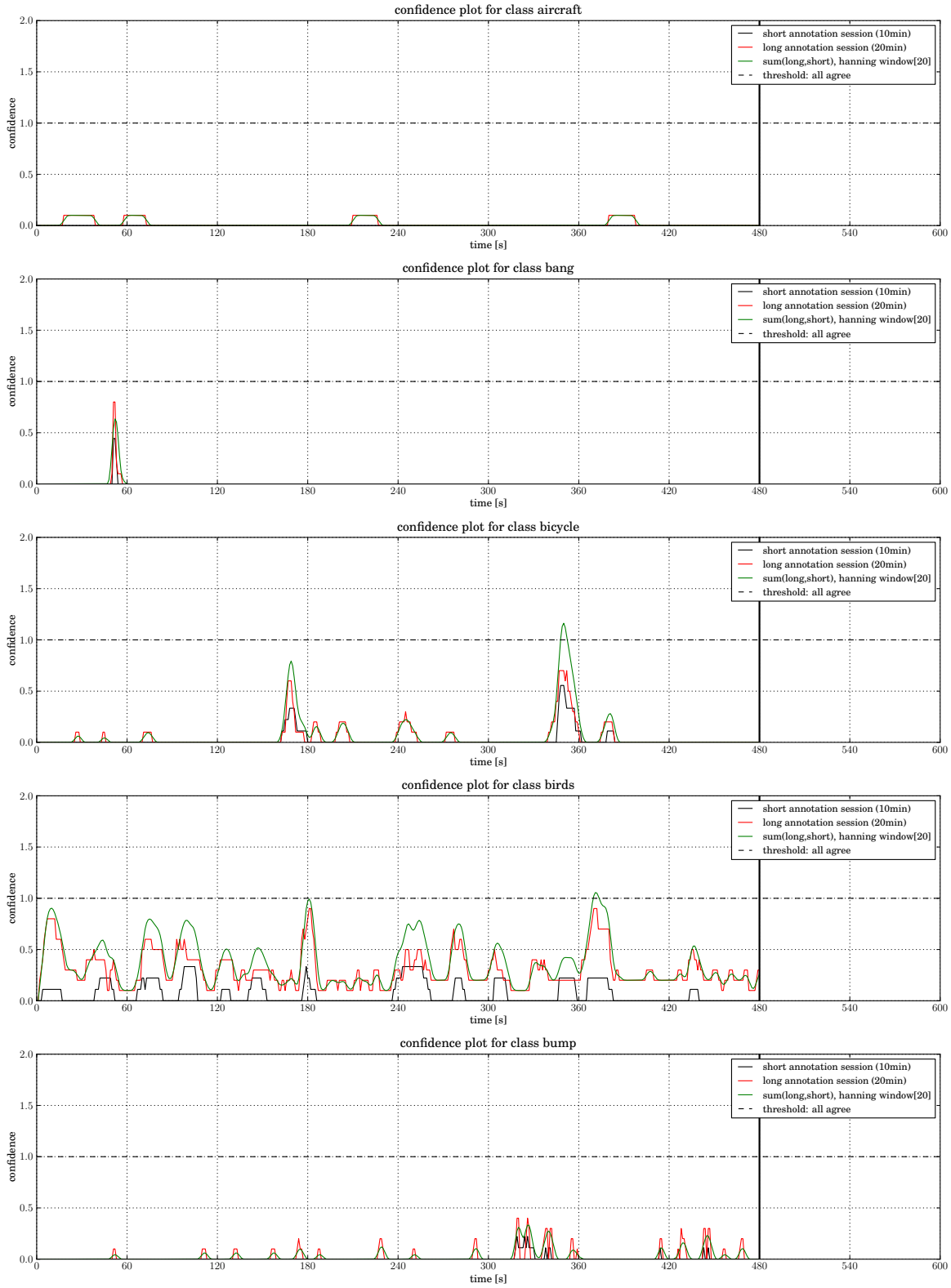

Figure 5.6: *Confidence plots for recording Part 1*

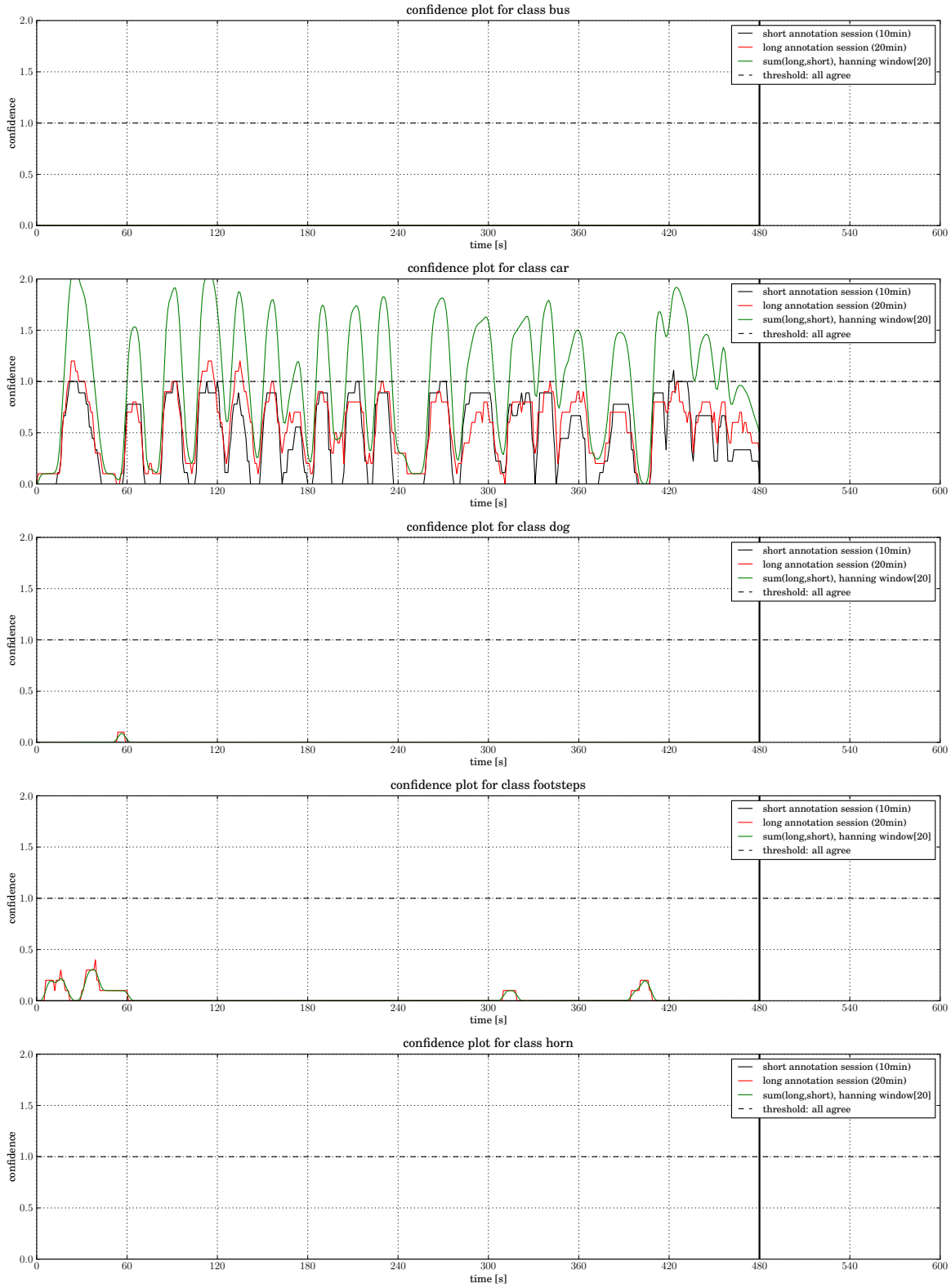

Figure 5.7: *Confidence plots for recording Part 1*

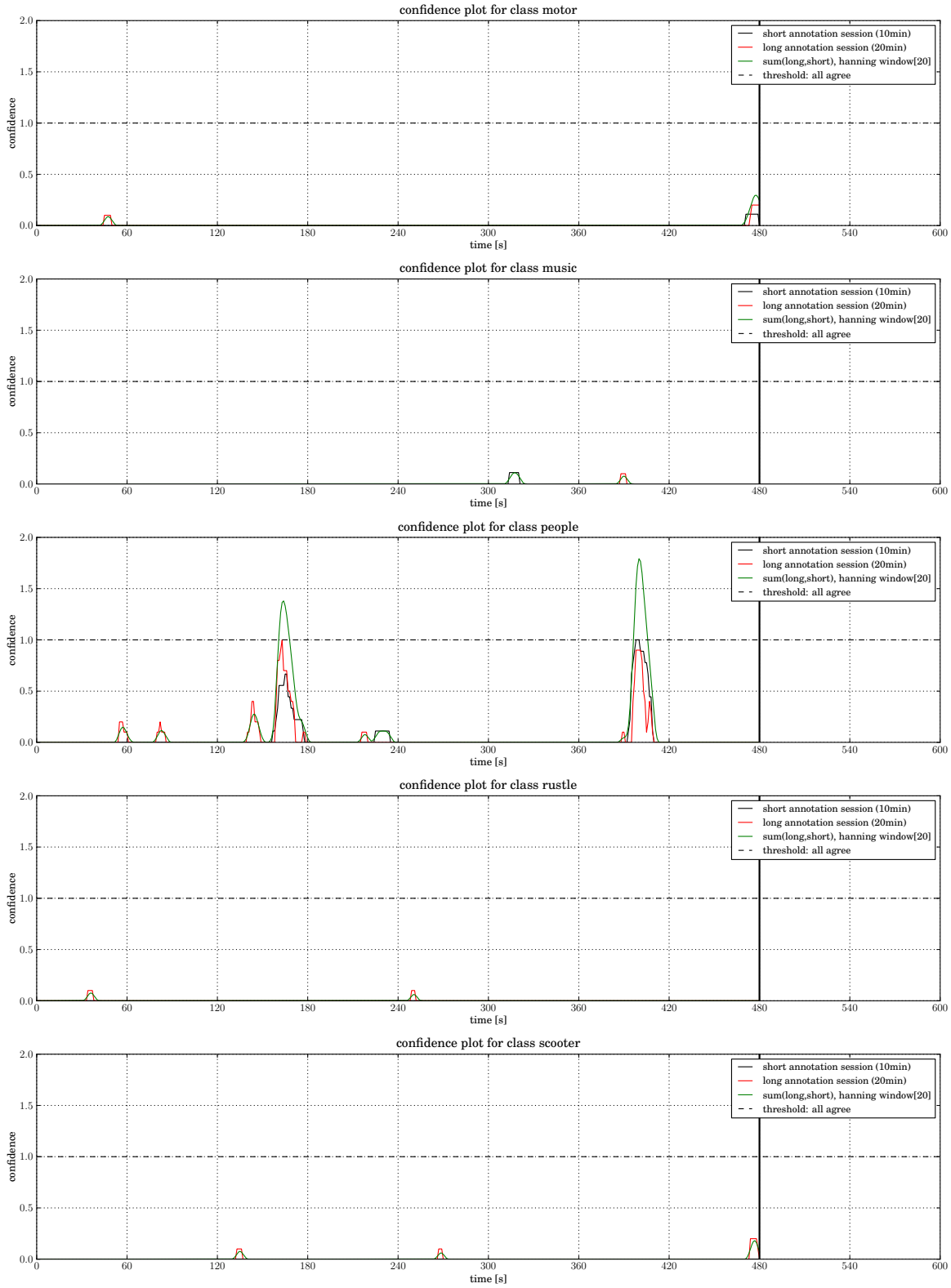

Figure 5.8: *Confidence plots for recording Part 1*

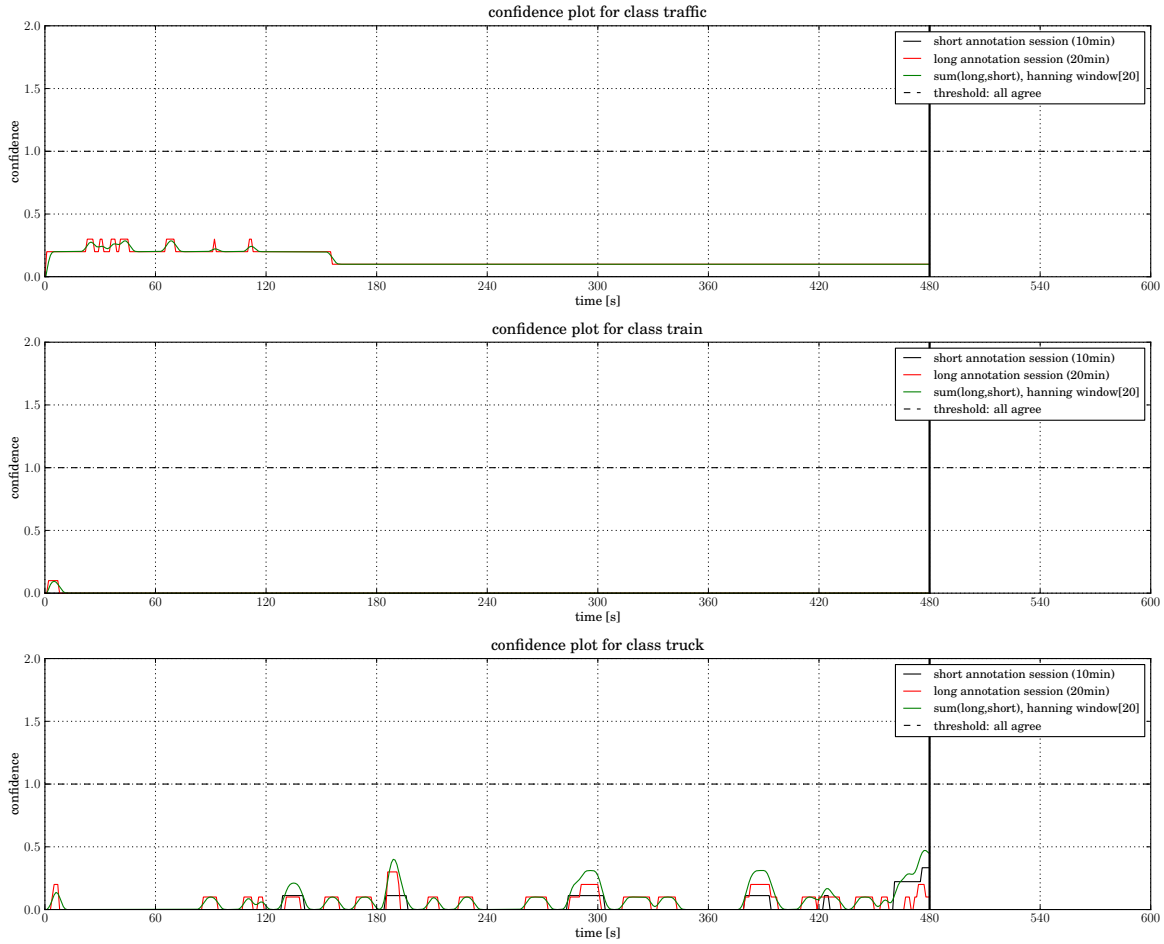

Figure 5.9: *Confidence plots for recording Part 1*

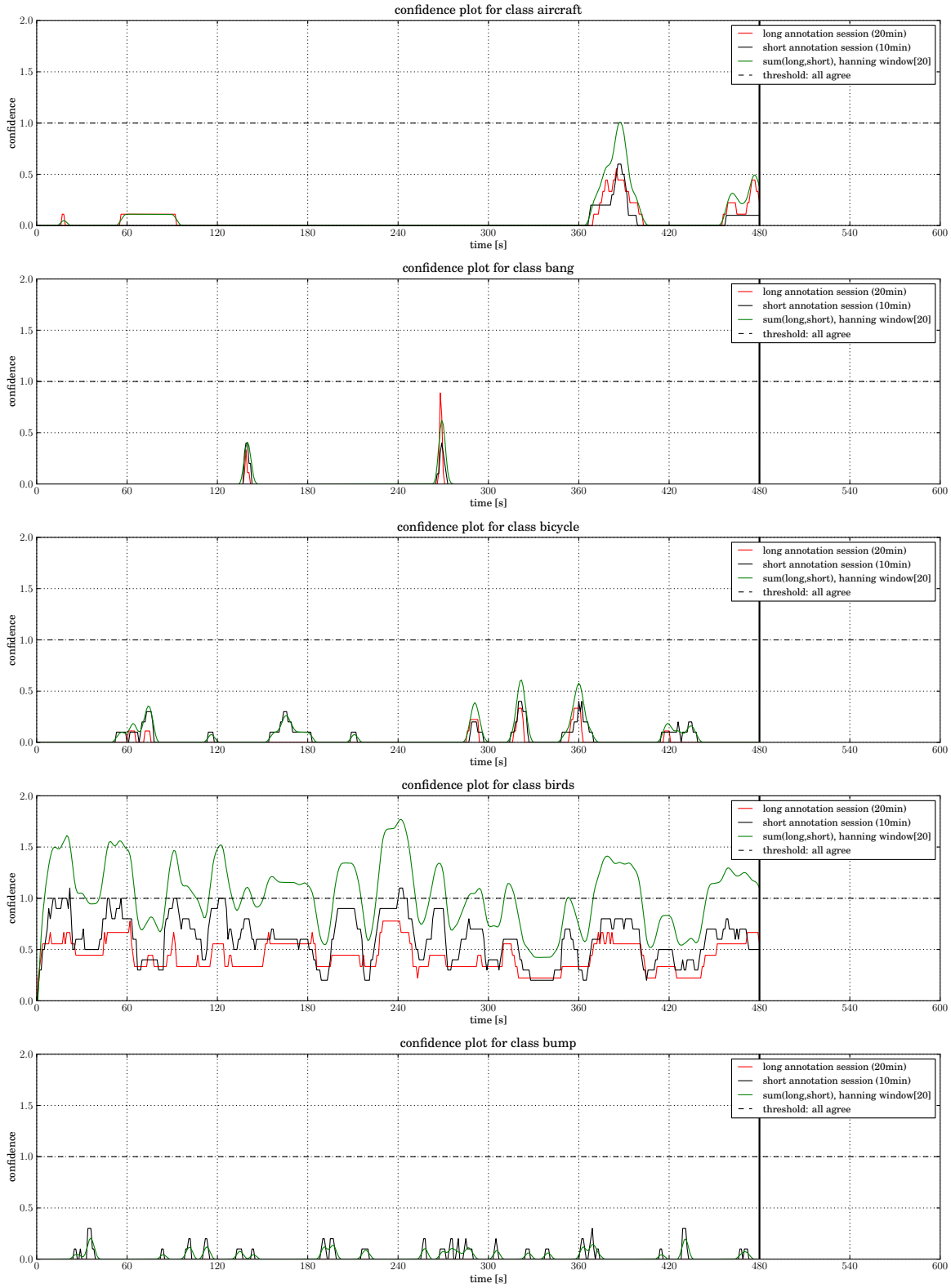

Figure 5.10: *Confidence plots for recording Part 2*

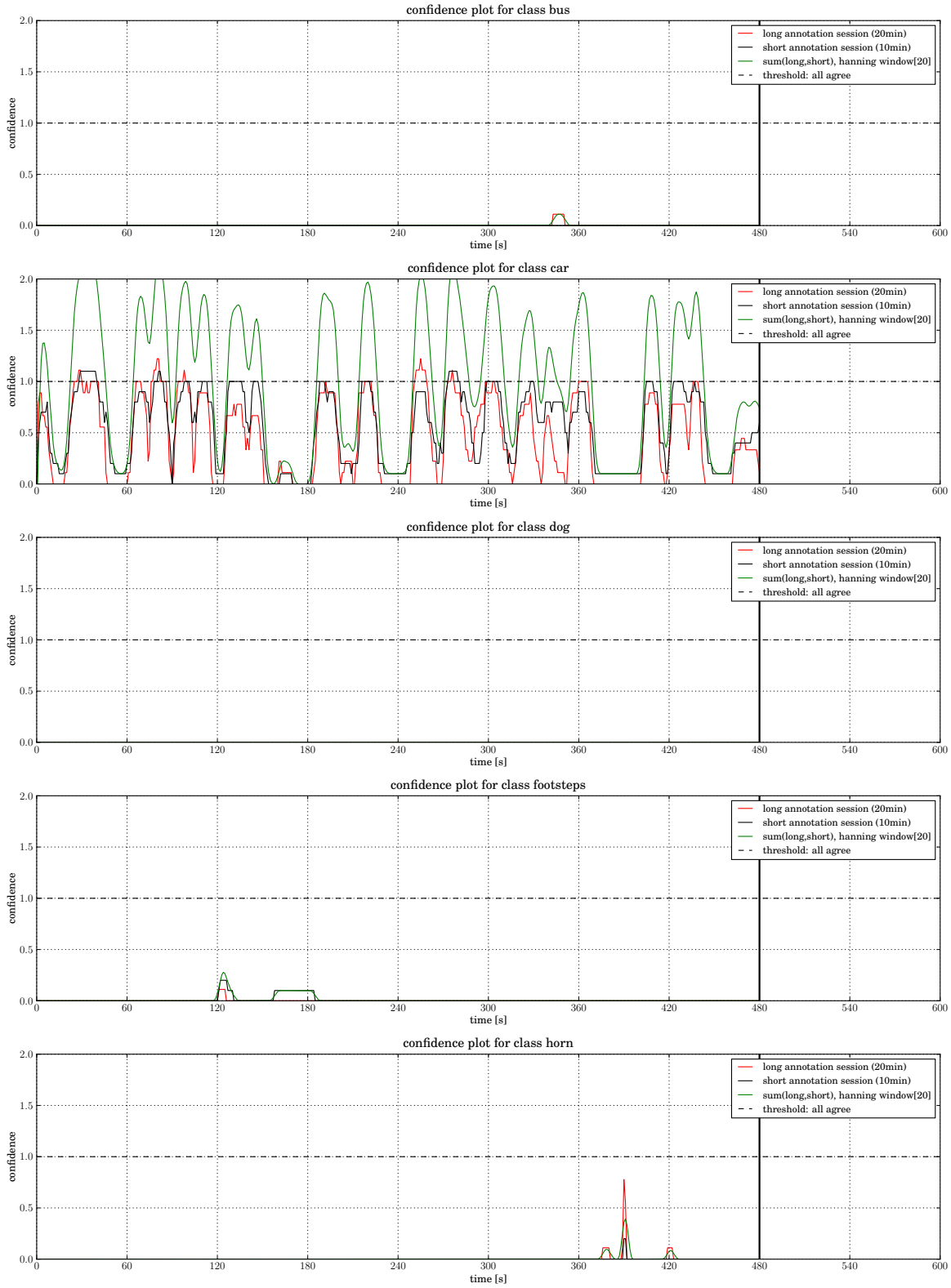

Figure 5.11: *Confidence plots for recording Part 2*

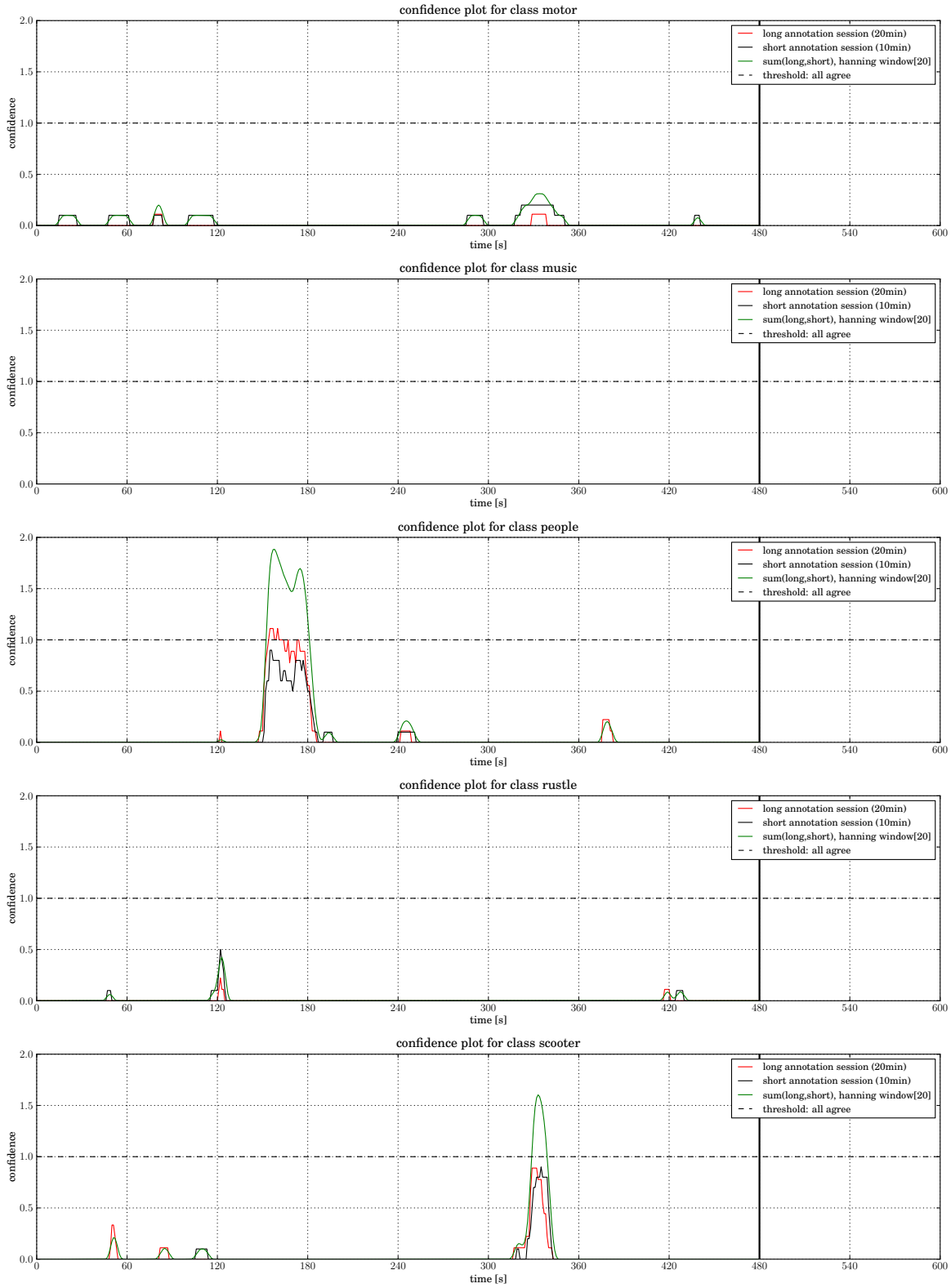

Figure 5.12: *Confidence plots for recording Part 2*

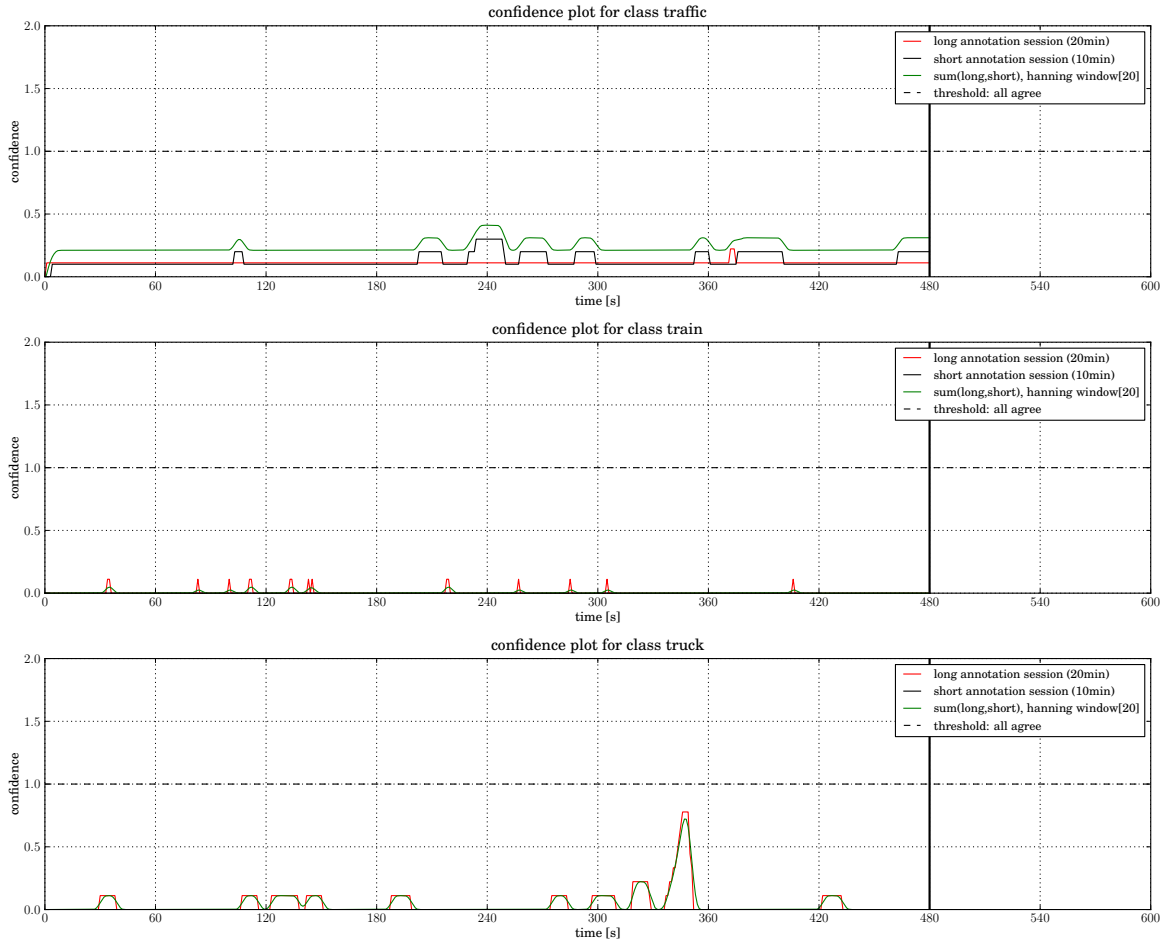

Figure 5.13: *Confidence plots for recording Part 2*

5.2.7 F-measures for each class

The mean F-measure or F-score (Hripcsak and Rothschild 2005) is calculated as the harmonic mean of the precision and the recall scores. This measure from signal detection theory is applicable in situations where the number of negative cases is large or unknown. This is the case in the current experiment: many sound sources can only be observed a few times in the recording.

The F-score was calculated from the precision P and recall score R; *precision* is a measure for the fraction of time an annotator was correct, and *recall* is a measure for the fraction of detections (of the occurrence of a source for that class) that a annotator annotated correctly. The definition of these three measures is then as follows:

$$P = \frac{TP}{TP + FP} \tag{5.1a}$$

$$R = \frac{TP}{TP + FN} \tag{5.1b}$$

$$F = 2 * \frac{P * R}{P + R} \tag{5.1c}$$

where TP is the true positive rate, FP is the false positive rate, and FN is the false negative rate.

Ground truth: thresholded confidence 'signal

To calculate the F-score, a ground truth (or *golden standard*) is needed. This was obtained by applying a threshold to the confidence 'signal'. Krijnders (2010) chooses a threshold of 30 percent: when 30 percent of the annotators agree on the presence of a sound source, this is adopted as the 'correct' annotation according to a set of annotators. This threshold is also used here. For both recording the threshold is applied to the annotations made in the 20 minutes annotation sessions.

The mean F-measures for each class is reported in table 5.4.

5.2.8 Correlation between confidence plots

To examine how the confidence plots (i.e. the combined annotations of all participants for a class) are related between the two annotation conditions (10 minutes or 20 minutes), the correlation was calculated for each recording $\{part1, part2\}$ and each class. Table 5.5 lists the cross-correlations between the two 'signals' that were derived from the annotations in each condition.
The entry `'-'` indicates that one of the signals was constantly zero and no cross-correlation could be calculated.

Class	Mean	Std. dev.
aircraft	0.00	± 0.00
bang	0.56	± 0.38
bicycle	0.46	± 0.32
birds	0.44	± 0.20
bump	0.13	± 0.18
bus	0.00	± 0.00
car	0.73	± 0.26
dog	0.00	± 0.00
footsteps	0.005	± 0.11
horn	0.00	± 0.00
motor	0.00	± 0.00
music	0.00	± 0.00
people	0.65	± 0.18
rustle	0.00	± 0.00
scooter	0.00	± 0.00
traffic	0.00	± 0.00
train	0.00	± 0.00
truck	0.00	± 0.00

Class	Mean	Std. dev.
aircraft	0.19	± 0.20
bang	0.32	± 0.24
bicycle	0.20	± 0.25
birds	0.50	± 0.38
bump	0.00	± 0.00
bus	0.00	± 0.00
car	0.74	± 0.10
dog	0.00	± 0.00
footsteps	0.00	± 0.00
horn	0.00	± 0.00
motor	0.00	± 0.00
music	0.00	± 0.00
people	0.80	± 0.14
rustle	0.16	± 0.32
scooter	0.59	± 0.26
traffic	0.00	± 0.00
train	0.00	± 0.00
truck	0.00	± 0.00

Table 5.4: *Mean and standard deviation for F-measures for each class. The left table corresponds to recording 1, the right table to recording 2.*

Class	**Recording 1**	**Recording 2**
aircraft	-	0.80
bang	0.97	0.72
bicycle	0.85	0.59
birds	0.53	0.67
bump	0.47	-
bus	-	-
car	0.87	0.87
dog	-	-
footsteps	-	0.55
horn	-	0.93
motor	0.66	0.48
music	-0.01	-
people	0.84	0.97
rustle	-	0.70
scooter	-	0.87
traffic	-	0.02
train	-	-
truck	0.41	-

Table 5.5: *Correlation values for the 'signals' of condition A and condition B for each recording. The correlation is undefined when one of the signals equals zero for all time intervals.*

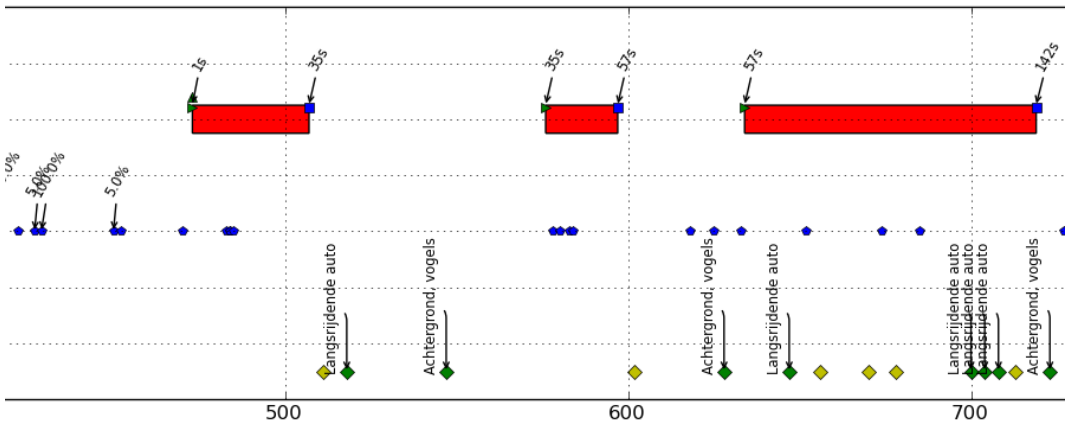

Figure 5.14: *Example fragment of a time line that was derived from the log file of one run. The red line on top indicates the time intervals that the audio player was running; time information is also provided. The blue dots in the middle indicate zoom actions. The diamond-shaped icons indicate the addition of annotations: green diamonds are annotations that persisted, yellow annotations were deleted later and were not present in the final annotations.*

5.3 Results: Participant behavior

As outlined above, an abstract representation of annotator behaviour was stored in the user data log files. This section first describes a method for visualizing the most important features of this data. These files also allow the extraction of statistics: numerical summaries of the data are presented in section 5.3.2. In order to be able to answer the research questions from section 1.1 on annotator behaviour, numerical summaries probably do not suffice to describe patterns in the behaviour of participants; qualitative descriptions can be of good help. These descriptions are presented in the discussion chapter.

5.3.1 Visualizing annotator behavior

An analysis script (implemented in Python) was developed to extract information from the log files and build up a *time line plot*. This plot shows graphical representations of the most important user operations and events that were initiated by the annotator in annotation application during a trial. The plot allows the researcher to inspect the annotator's behavior at a glance. These time line plots were created for each participant and each trial. In figure 5.14 a fragment of such a graph is presented as an example.

Using the registration instruments described in section 4.2.2 the following numerical data were extracted or calculated:

Audio player actions: The number of times the audio player was started, stopped or paused the audio player. This results in a number of fragments in which the audio recording was divided according to the play-and-pause behavior.

Zoom actions: Subjects were free to zoom the cochleogram representation in and out (on the time axis, the frequency axis was fixed to allow identification of sonic information

in the frequency plane). A special zoom action is the 'shift' in which the cochleogram window is shifted to backward or forward in time. Not all subjects used this feature, therefore the table mentions the occurrence of **shifts**.

The number of annotations created by the subject. Not all annotations persisted during the experiment, some were deleted: the number of deleted annotations is also mentioned.

Duration of the total trial.

5.3.2 Quantitative analysis: event frequencies

Two indicative statistics were calculated for each trial and compared between conditions: the number of *play intervals* and the number of *zoom actions*. The means are presented in table 5.6 and plotted together with the standard deviation in figure 5.15.

Condition/Recording		Play intervals	Zoom actions
Recording part 1	Short (10 min)	21.56	55.89
	Long (20 min)	80.50	64.10
Recording part 2	Short (10 min)	13.60	42.10
	Long (20 min)	71.44	111.33

Table 5.6: *User data from the experiment for each recording and condition: the mean number of play intervals that indicates the number of fragments in which the annotator listened to the recording, and mean number of zoom actions, number of times the annotator changed the cochleogram view by zooming in, zooming out, or shifting the window to the left or right.*

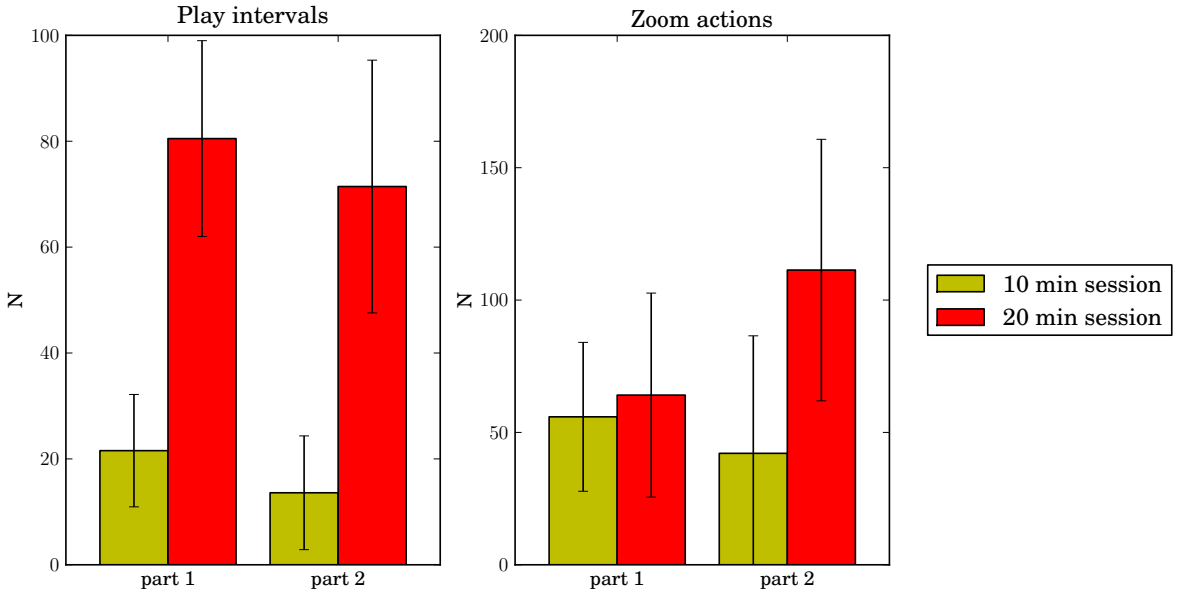

Figure 5.15: *Histogram plot of the data from table 5.6. Please note that the scales on the x-axis differ for the two plots.*

5.4 Survey results

The results for the survey (see section 4.2.3) are discussed in the next chapter, in section 6.1.

Chapter 6

Discussion

The previous chapter presented the numerical and graphical results for an experiment that tested subject behavior in an annotation task using a specialized software tool. This chapter discusses these results and answers the following questions:

1. Regarding the *annotations*: What patterns can one observed in the annotations that resulted from the experimental annotation sessions as described in chapter 4?

2. Regarding the *user behaviour data*: what do these data tell about the *strategies* that the participants employed during the session?

The following sections discuss the results in the same order as they were presented in the previous chapter.

6.0.1 Annotations

Section 5.2 presents the number of annotations that the subjects entered for each session. Figure 5.1 reveals that in condition A, where subjects were presented with a short 10 minutes annotation session followed by a longer 20 minutes annotation session, most annotators enter about twice the number of annotations in the 20 minutes session; hence their number of annotations per minute is relatively the constant. For condition B the picture is slightly more complex: most subjects make less annotations under the condition with increased time pressure, but in four annotators we observe the opposite: these annotators even entered more annotations in the second 10 minutes session. This most likely reflects the difficulty of some annotators with becoming aquainted with the annotation task in the first session; after some practice their performance has increased.

We will start the discussion of the results with section 5.2 where we present the number of annotations that the subjects entered for each session. As expected, a difference is observed between annotation frequencies for the two conditions. Figure 5.1 reveals that in condition A, where a short annotation session precedes a long annotation session, most annotators enter about twice the number of annotations in the 20 minutes session; hence their number of annotations per minute remains constant. For condition B the picture is slightly more complex: most subjects make less annotations under the higher time pressure, but four annotators even made more annotations in the second 10 minutes session. This may reflect the difficulty of some annotators with becoming acquainted with the annotation task in the first session; after some practice their performance may have increased significantly. It may also reflect the tendency of some annotators to be more precise in the second session if they

are given extra time for a recording of the same length; the extra effort they took then is reflected in *better* annotations, but not in the number of annotations added.

Figure 5.2 presents average annotation counts per annotation session (on both parts of the recording). In the first session the annotators from both conditions produce on average the same amount of annotations per (trial) minute; however in the second session the number of annotations for each condition is on average comparable.

We explain these findings by hypotising that the time constraints in first session determine how annotators perform in the second session: annotators that are under time pressure in the first session do not 'relax' in the second session, but on average produce about the same number of annotations per minute. In contrast, annotators who are under more time pressure in the second session aim to produce the same level of detail as they did in the first session. The training effect (annotators becoming acquainted with the task) may also have influenced these results.

This figure also shows the average number of deleted annotations; in the short annotation session almost no annotations are removed, probably because the for this the annotator has to stop the recording and decides that this cost is not worth the effort.

Class naming and frequencies

The annotations reveal a wide variety in the classes used to describe the sound sources; in total the annotators created 189 unique classes (uppercase characters were set to lowercase) during the experiments. Comparison of so many different classes between annotators is futile, therefore 18 'common' classes were selected and each class was mapped by hand to one of these 'common' classes.

Plot 5.3 visualizes the combined results for the first annotation session by presenting the mean and standard deviation for each 'common' class. It reveals that the most popular class description is 'cars': in both the 10 minutes and 20 minutes condition the majority of the annotations indicates classes mapped to 'car', with not much differences between the conditions. However, the variation of the number of annotations in class 'car' is much higher for the longer annotation session.

Another popular class is 'birds', but here differences between the 10 minutes and 20 minutes condition are observed: three times as much annotations are assigned class 'birds' when annotation time is less constrained. The plot also indicates that in the longer annotation session the annotators are more specific in their annotations and more often choose classes indicating the presence of trucks, scooters and traffic in general in the recording. The results for the class 'bump' indicates that the bumps in the road (close to the microphone) are more often annotated as a separate sound source in the long annotation session, but again the variance between subjects is high; not every annotator included this class, while others annotated each car crossing the bump.

The picture is different for the second annotation session (see figure 5.4): the class 'car' is still prominent, but the increased variance is not observerd for the longer annotation session. In-

terestingly, the class 'birds' shows a higher mean annotation count for the *short* session (but still with high variance); on average the extra time available does not cause annotators to add more annotations of 'birds'. Interestingly, some annotators perceived a 'train' in the recording, while there is no railroad in the vicinity of the recording location. The small plane that circled the town of Assen during the second part of the recording is perceived and annotated in both conditions.

Combined annotations and confidence plots

Section 5.2.6 describes how the annotations from all participants were combined to create a plot that shows the number of annotations for each 'common' class on a time axis. The number of annotators making an annotation for a class can be seen as a 'confidence' measure that reveals how likely it was that a sound source of that class was present, for each time segment. This 'confidence' can also be described as a combination of saliency and recognizability of a sound source. When divided by the number of annotators this annotation count can also be seen as a measure of agreement among annotators on the existence of a sound source for that period.
These segments were here set to 60 seconds and plots were created for each class. Included in one figure are the plots for the 10 minutes (short) annotation session and 20 minutes (long) annotation session, and a plot of the 'sum' of those plots, in black, red and green respectively. This sum is smoothed to correct for the rigid changes that are caused by the all-or-nothing boundaries of the annotations. To improve discriminability the sum is calculated as the direct, non-weighed sum.

The plots are clearly different for each class, therefore we will discuss most classes in isolation here. Some classes show between-class interference that make it worthwhile to discuss the corresponding plots together.

Traffic sounds are most prominent: Beginning with the class that was annotated most, namely 'car', one can see that both the signals for 'car' peak above 1.0 several times; the resulting summed plot also peaks above 2.0. This indicates that for some intervals (almost?) all participants annotated sound sources that can be described as cars (but this is not necessary; see the remarks below). The plot for 'car' also indicates clear boundaries for the presence of car sounds; apparently car sounds cause all annotators to either make an annotation for that class, and annotators seldom hear a car when the other annotators don't.

Birds are annotated when other salient sources are absent: The plot for recording 2 reveals that there were two relatively long periods where, according on the annotations, no cars were audible; these periods can be observed around 150 and 380 seconds into the recording. When this observation is combined with the plot for 'people' we expect that in the absence of traffic sounds the annotators attended to sounds indicating human prescence and activity; this is indeed the case.

For the second 'silent' period the attention seems to be directed to birds, because the plot of this class shows a clear peak for the corresponding period.

Salient sound sources result in short peaks: One can observe from the figures that salient, sudden sounds result in sharp, high peaks in the plots. This effect shows for the class 'bang' that holds all class descriptions that correspond to sudden, explosive acoustic events, such as 'knal', 'boem', 'door' and 'autodeur slaat dicht'. Apparently these events are salient enough to appear in the annotations and are annotated by a large proportion of the participants.

Interference between related classes: Another difference that shows when comparing different plots is that for the classes 'scooter' and 'truck'. Both relate to traffic, but the class 'truck' is used significantly more in the 20 minutes annotation sessions - this holds for both recordings. This effect is not observable for 'scooter': apparently a noisy scooter passed in the second recording, around 320 seconds into the recording, but this was annotated as a separate class about as many times in the short as it was in the long annotation session. We hypothesize that a truck may be annotated as the already present class 'cars' (because many cars have passed before), but a scooter does not fit under that class description, hence annotators decide to make a new class for this source. If this hypothesis is correct, this illustrates how previous sonic events shape the class description list that in turn shapes the annotations.

Annotations of 'bumps': The classes that fit under the description 'bump' are also worth closer inspection. This class shows the same sharp peaks as does the class 'bang', but only a limited number of annotators annotated the 'sound of a car meeting a bump in the road' specifically. For the first recording the 'bump' sound was annotated mostly in the 20 minutes annotation condition, but in the second recording it was only annotated in the 10 minutes; one might expect the opposite, as the bumps can be regarded as extra details to the car sound for which the annotator is not expected to make annotations under time pressure. A possible explanation is that listeners do not pay attention to details on this level when put under time pressure, and thus will only detect them in the first recording when they have enough time to listen carefully. Annotators then persist in their understanding of the task for the second session and may regard the 'bump' sounds as irrelevant, or may completely omit those sounds.

Human sounds attract attention: Most annotators clearly distinguish human sounds, more in the long session than in the 10 minutes session, but still present in the latter. Apparently there were some occasions of human speech close to the recording setup, as the plots for both recordings show clear peaks.

The plots suggest some interference of different classes in the long annotation session that is absent under the 10 minutes time constraint.

In the second recording the sound of a heavy truck can be heard, around 300 seconds into the recording and around 390 seconds into the recording. The plots for this class reveal that some annotators assigned the class truck to this sound source, as a clear

peak for those two timepoints can be observed.

Footsteps however were not that important to the annotators, even while these sounds were clearly audible. Only some participants annotated these sounds.

Some non existent sound sources were also annotated; the class 'train' reflects how one annotator repeatedly heard a train passing, while there is none. Also the annotations for aircrafts occur much more often in the annotations than the real aircraft (a small plane flying above Assen) is audible in the recording. A possible explanation for these 'confabulations' in the case of aircrafts is that these observations are hard to check against further evidence from the environment: the presence of an aircraft does not interfere or conflict with the other sound sources, and these observations of other sources therefore do not approve nor disprove the reliability of the first observation. Hence the hypothesis that there was an aircraft will persist, and listeners will be likely to hear one, even if it was not there.

Confidence plots: discussion of method

The method described above provides a good summary of the annotations from different participants. There are however some potential problems with this approach:

1. **The choice of 'common' classes is arbitrary;** different common classes result in quite different plots. Applying a (hierarchical) ontology may solve this.
2. **Annotations may overlap:** when an annotator adds two annotations ascribed to two classes and both these classes are mapped to the same 'common' class, one annotator attributes double to the sum of annotation. This can be corrected by checking the overlap in annations from one participant, but this results in loss of information.
3. **Some annotators added annotations for very long time regions** indicating that the sources was 'always present'; in this approach, this causes the 'signal' to never become zero and the whole plot to shift upwards. It is not desired, but deleting these long annotations cause information to be lost. This can cause the 'confidence' rate to be higher than 1.0. This 'uplifting phenomenon' also can cause the curve corresponding to the 10 minutes session to peak above the 20 minutes session signal.

Thresholded 'confidence' as a ground thruth

Table 5.4 presents the F-scores for each class, for both recordings; the definition was provided allready in section 5.2.7. A high F-measure for a single annotator indicates a close match between that participant's 'sound source' detection and the sound source annotations that most annotators agree about. Mean F-scores can be observed to be high (> 0.5) for classes $\{'bang', 'car', people\}$. F-scores differ between the two recordings for most classes; large differences can be observed for $'aircraft'$ and $'scooter'$, reflecting the presence of prominent sounds of that class for one recording and abscense for the other.

Comparison to earlier results

These results can be compared to (Krijnders 2010) who describes the variation in F-scores in a live annotation experiment. The current experiment uses two recordings that were made consecutively, (in practice: split afterwards) during the experiment described in Krijnders (table 1, *'location 1'*). The class lists do not match completely, but for the most prominent sound sources the results can be compared. We observe that the mean F-scores per class are much higher for the current experiment: we hypothize that the *off-line* annotation setting allows annotators to detect much more different sound sources and allows them to be more precise in indicating the appearance and disappearance of a source. Standard deviations are however much higher for the current experiment, indicating large inter-annotator variation in annotations.

6.0.2 Annotations: Qualitative analysis

The previous sections gave a quantitative approach to the annotations was described. One can also look at annotations from a qualitative perspective: when inspecting the sets of annotations, what can be concluded about the strategies and priorities of the annotators? How do the annotations differ between participants?

1. Not all subjects annotate the same sound events, some events that were annotated by one subject were dismissed by another.

2. The 'granularity', i.e. the level of detail of the annotations varies among subjects. What is annotated as one sound event by one annotator may be annotated as two or more events by another subject.

3. The number of classes a subject used to annotate the soundscape recording differs; subjects were free to add or remove classes at will. This also introduced semantic differences in the annotations: what is described as a 'car' in one annotation set may be described as 'traffic' in another, but the tool currently provides no means to represent the hierarchical relationship between these descriptions, see 5.2.2 for a description of the class lists. There is much room for future research here: see section 6.2 for a discussion of the possible routes.

6.1 Subjective experience: surveys

This section discusses the results from the survey that was held among the participants; it was described in section 4.2.3. From a total of 21 participants, 17 filled out a survey that questions their experience in carrying out the annotation task.

6.1.1 Subject's report of their strategy

Q:Please describe your strategy in carrying out the annotation task.

This question resulted in a wide variety on answers. Some respondents also describe why their approach was not fruitful; one participant describes how he repeatedly listened to the same part of the recording (in the 20min condition) to identify more sound sources, but eventually found himself taking too much time to complete the task. About half of the participants report that they tried to avoid to stop the recording in order to make an annotation; rather they tried to let the sound play and add annotations meanwhile.
Participants report different ways of using the zoom function: some report to set the zoom level to a small portion of the signal and leave it like that for the rest of the trial, sliding it back and forth. Other participants report to zoom in once they hear a sound, to enter an annotation and zoom out again. A few participants motivate their limited use of the zoom function by pointing at the time constraint - in their view there was no time to look closer to the spectogram.
A particant reports a special technique to be able to annotate quickly: he describes how he clicks and drags as soon as he hears a new sound source, then releases the mouse after the sound is gone and immediately selects a class label.
The participants do not report much details on their strategy in selecting a class for the annotated time regions: only one participant notes that she first listened to identify sounds that could constitute a class, and later she focussed more on the recognition of the classes allready created.
One subject reports a change in strategy when he finished listening to the recording and started it again from the beginning; he reports focussing on the 'blue' (silent) areas of the spectogram to identify less prominent sound sources.

6.1.2 Subject's report on the reliability of their annotations

Q: Did you carry out the task with care?

All but one subject report that they made the annotations with care; the remaining participant reports that he worked with 'moderate' care. Some subjects report that in the 10 minutes annotation session they were forced to work less precise. One subject reports that in the second session her annotations were more reliable because of the extra instructions that she received during the break between the two sessions.
Three subjects report that their annotations might be constituted too carefull; they concluded that they could have done better in dividing their time over the recording.

6.1.3 Subject's report on the annotation tool

Q: Did the software work well? What is good and what could be better about the software? (Please list both)

Subjects were asked to report strong and weak points of the software. **Strong points** of the annotation tool are, according to the respondents:

- The tool is easy to understand (7 mentions)

- The zooming is useful (1 mention)
- Visual representation of signal is helpful (1 mention)
- Making an annotation is easy and the program works well (1 mention)

Participants in the survey identified the following as **weaker points** of the annotation tool:

- The time indicator cannot be moved by hand (only by zooming) (3 mentions)
- There is a bug in zooming: screen turns black when selecting the beginning of the recording (2 mentions)
- It would be useful if the cochleagram window moves with the audio playback position (1 mention)
- Annotating from right to left does not work (2 mentions)
- It is impossible to rename or delete categories from the list (1 mention)
- The bars representing annotations (time intervals) are small - hard to select (1 mention)
- The playback buttons are not intuitive (2 mentions)
- Space bar key does not start/stop playback (2 mentions)
- A zoom level slider would be useful (2 mentions)

6.1.4 Subject's report on their perception of the environment

> *Q: How sure are you that the sound sources you recognized in the recording were really there?*

Subjects were asked to report how sure they are about the sources that they recognized - is there a chance that they might have mistaken sound sources for others?
From the respondents most participants reply that they have confidence in their reports. Two (Dutch) respondents reply that the recording was realistic - this might imply that they misunderstood the question and read it as 'do you think you were fooled during the experiment?'. Three participants reply that it is hard to hear within-class differences, for example the difference between a heavy car and a truck. Two participants note that purely based on audio multiple interpretations of the recording are possible. One participant states that listening 'live' (in-place) would provide better results; another participant replies that audio-based recognition is not reliable. Concerning the completeness of the annotations, one participant responds that she thinks it is more likely she missed sound sources than that the annotations are wrong.

6.2 Future work

This thesis describes the development of an annotation tool, a method to use it for gathering soundscape annotations and an experiment that tests this method and tool. Based on the results of the experiment it was argued that the tool is successful in allowing annotators to comfortably and efficiently annotate a lengthy soundscape recording. The development of this lightweight annotation tool provides a good starting point for further research in soundscape annotation and audition in general. This section provides an overview of further possibilities in research in soundscape perception and annotation in general, and particularly using this tool.

6.2.1 Use different soundscape recording and reproduction methods: take ecological validity of soundscape reproduction into account

Currently single-channel recordings are used for annotation. These signals are presented on both channels of high-end headphones. There are however more ways to present soundscape recordings to the annotator: one possibility is using stereo speakers. Another option is the usage of multi-channel recordings that of course need to be presented using multi-speaker systems. (Guastavino et al. 2005) demonstrated that different soundscape reproduction methods invoke different reporting modes in humans when describing the percieved soundscape.The experiments of Guastavino et al. were limited to the subject's report of their acoustic environment; experiments testing this factor in a sound event annotation setting need to be set up to determine the influence of the reproduction method on the resulting annotations.
In (Krijnders 2010) the difference in annotating a soundscape 'live' (when actually present) and annotating a recording played over (classroom grade) loudspeakers is tested. In Krijnder's pilot experiment no significant differences were found.

6.2.2 Adding context information to the system

For the annotator:

In the current approach the annotator's task is to recognize sound sources without any form of context - only a very shallow description of the surroundings of the microphone was given, such as 'the recording was made outside a large building alongside a road in a small town in the Netherlands'. The discussion of audition presented in section 2.3 argued that top-down, knowledge driven processes are important to disambiguate complex stimuli. Anecdotal evidence also indicates that for humans it is very difficult to recognize a sound source outside its normal context. Therefore one can argue that in the case of annotation a soundscape recording a certain amount of context information needs to be present to allow reliable recognition of sound sources. Different levels of context information can be tested in a human annotation experiment to reveal the effects on the resulting annotations.

6.2.3 Provide more visual information to the annotator

In the current annotation tool the annotators are provided with a cochleagram representation; in addition to that a waveform representation may be presented. The cochleogram panel could also be enriched by adding graphical representations of physical features, or a proposal for a possible segmentation.

6.2.4 Test different cochleogram representations

In section 3.1.1 it was motivated why in this project annotators are provided with a cochleogram image: it allows subjects to identify similar acoustic events by sight and provides a reference point when seeking through a long recording, for example when the subject decides to inspect a portion of the recording for a second time. The process of deriving this cochleogram representation is provided in (Krijnders and Andringa 2009b). However, using a cochleogram is somewhat arbitrary: a spectrogram may contain similar information, and even a waveform plot may allow some guidance and recognition by sight. In future work the different options for presenting a visual representation can be tested as an experimental condition in an annotation task.

6.2.5 Assess the usability of the tool

Annotating sound recordings is time-consuming, and in practice an annotation tool will be used for many hours. It is therefore worthwile to look for possibilities to improve the usability of the tool. Applying knowledge from human factors research will allow the developer to identify potential problems in the (cognitive) ergonomics of the application.

6.2.6 Introduce ontologies

In the current approach annotators are free to choose an 'ontology' themselves: no constraints are imposed on the labels that describe the selected annotation segments. This freedom in choice of class labels produces great variation in the descriptions that different annotators attach to the annotation segments. For some applications of annotations this variation may not introduce a problem, but it certainly is problematic when the annotations are used to train and test an automatic recognizer. In this case uniform descriptions are needed to allow machine learning algorithms to learn patterns in the provided data. Different sets of annotations can be merged by assigning matching classes manually, but it is then not clear if the resulting annotations still hold all the information present in the original labelling.
The introduction of an ontology may solve both problems. The annotator can select from an ontology which description best matches the sound source that he or she identified in the signal and the ontology itself limits the variation in chosen labels. However, the choice of the ontology is important as it also limits the freedom of the annotator to express in a precise manner what sound source exactly was present, limiting the value of the annotations. We suggest that for each application of annotations this balance between freedom and generality to be chosen carefully. A possible option is to use Wordnet to obtain semantical information

from Wordnet (Fellbaum et al. 2001).

6.2.7 Let the tool compensate for unwanted attentional phenomena

We speculate that when the accuracy of automatic sound recognition techniques increases, these techniques will detect sounds sources from the signal that a human annotator would have missed due to the attention and omission effects described in the Background chapter.

6.2.8 Implement assisted/automatic annotation

Manual annotation of a soundscape recording by a human annotator is time-consuming. An alternative method future research is to let an algorithm partly or completely carry out the annotation process.

Part of this was allready done, be it very preliminary: (Tzanetakis and Cook 2000b) describes a 'semi-automatic annotation' setup, in which te task mainly is to segment audio recordings. The results described in this conference paper are very preliminary, and the segmentation paradigm is rather limiting as overlapping sound sources cannot be annotated. The strategies employed by the subjects are however interesting: in the experiment subjects showed a wide variety in segmentation methods and different sound sources induced very different segmentation choices (for example: segmentating a song in verses and refrains is something different from segmentation a radio broadcast); this variety is also reported in the annotation experiment described in this thesis.

Current problems with this approach are listed in (Tzanetakis and Cook 2000b):

- Current systems are not perfect and make errors. The source of this error is likely to be twofold: it may result from the choice of features, or the choice of the learning algorithm; improvements on this part may solve the poor recognition scores. Current systems simply are not good enough to work in the real world.
- Human auditory perception shows prominent within-subject differences: listeners tend to differ in the reports of the sources that they identify in a soundscape. However, there are properties of the audio that all listeners agree on, and it might be enough for automatic annotation to just extract these attributes (and not mimic human performance).

Concluding, one can say that there is much room for future research on the best method of collecting annotations for each specific situation.

Chapter 7

Conclusions

This last chapter concludes this thesis by looking back at the previous chapters an summing up what was learned about the annotation task, and hearing and listening in general.

7.1 Conclusions

Semantic annotations are useful, in the first place to use as a ground truth to train automatic sound recognition systems upon.

The current state of automatic sound recognition techniques is unsatisfactory: in two fields, namely voice recognition and music recognition, advancements have been made. Any experienced computer user can use a voice recognition program to dictate a letter to his computer. Current solutions score above 95% in word recognition; not perfect, but it works. However, these solutions only work well under controlled conditions on a 'clean signal': typically a commercially available voice recognition package is not able to distinguish between to co-occurring speech signals and thus screws up completely if a co-worker enters the room.
What is needed are systems that can function in the real world, under complex circumstances and in the presence of sound sources that interfere with the target source.

Soundscape recordings used for research should be made in the real world, and should reflect the stimuli that one wants a recognizer to be able to work with. Therefore the recordings used in this thesis were made outside on the street, in a real-life setting. Not all databases of environmental sounds take this approach (and there is confusion about what exactly are environmental sounds).

There is not much work done on the full semantic annotation of real-world sound recordings. A few examples of sound recording databases were listed the Background chapter, but most contain 'clean' recordings. There are a few software tools that allow to make semantic annotations as used in this project; ProjectPad comes closest to our approach, but lacks a visual representation of the audio data.

Top-down processes are important in audition: Humans listen to sounds from sources in the environment through their complex auditory system, and their auditory perception is strongly influenced by higher-order cognitive processes. A group of processes generally called 'attention' is important in the selection of stimuli that reach awareness and will eventually be reported in an annotation setting. Attentional processes can

promote the selection of stimuli, but a strong attentional focus on stimuli can cause other stimuli to be neglected. Attentional processes probably function upon a sparse representation of the perceived stimuli: the 'gist'. Gist is a representation that contains all information from the (here: auditory) scene to allow selection of interesting or task-relevant regions for further examination. The concept of gist promises to explain how perceptional processes cope with the endless stream of sensory input.

To report allow a human listener to produce a structured report on the sound sources that can are perceived in a soundscape or recording, an annotation tool was developed. The current implementation presents the annotator with the sound and a graphical representation of the audio. The interface allows the annotator to indicate time regions and attach semantic class labels to those regions, thus creating a set of annotations. This tool has proven to be easily understood by novice users.

The experiment with 21 participants making annotations shows that the majority of annotators is well capable of producing reasonable annotations under time pressure. The annotators however vary in the number of annotations that they can make in the time available in the trial: the fastest annotators add twice as much annotations as the slowest participants. High variety is also observed in the sound sources that are annotated: initially the participants created too much different classes to allow comparison between subjects. The mapping upon 18 standard classes reveals large differences between classes: traffic sounds apparently were most prominent to the annotators, followed by animal and human sounds. When combining the annotations to obtain the number of annotations for each class, for each moment in time, it was shown how these combined annotations be interpreted as a confidence measure for the presence of that type of sound source in over time.

Listening is selective , and so is annotating a sound recording. The results indicate that salient sonic events are not guaranteed to be annotated by every annotator; some salient sound events do not even cause half of the annotators to make a corresponding annotation. This implies that time pressure may lead to the omission of salient stimuli in this task. Furthermore, the annotators seem to create a perception of the task and stick to that interpretation: when they have annotated a series of cars passing, they are likely to annotate another car instead of a co-occurring salient sonic event. To allow annotators to detect and annotate sound sources that are not closely related (for example annotate: annotate all human sounds *and* traffic sounds) it may be necessary to let the annotator do the task twice.

Annotator behavior in the test session also shows large differences between subjects. When the time constraint of the trial is loosened almost all annotators use that extra time to pause the audio playback and/or listen to interesting parts of the recording more than once. Also the zooming functionality is only used by part of the annotators; this may imply that the annotators *not* using it miss information that can be obtained through examining the visual representation.

7.2 General relevance of this research

This thesis has relevance for several fields and types of research, at least including the following areas:

1. Sound is everywhere: as result of the ever increasing urbanization 'sound management' is becoming ever important, especially in preventing noise pollution. Physical measurements may not provide enough information to predict and control human auditory perception; **annotations to a soundscape provide a structured report of the human perceptual experience**. This thesis provides a method to obtain those annotations.

2. Sound recognition systems have high potential. Environmental monitoring and security applications can greatly benefit from the development of robust, real world sound recognizers. Robots will inevitably fill up our private and public space in the future and can benefit from techniques that allow them to monitor and analyze their surroundings through sound, just as sound is important to animals and humans. When robots carry out human tasks, one can assume that a human-like sound perception allows the robot to perform his tasks. **The human annotations from this thesis may provide a basis to develop such a system for robots**.

3. With the rise and availability of digital information systems most people become used to the possibility of scanning and searching large databases of digitalized information. For text many successful search techniques have been developed, but for sound these are still under development. *Human annotations may provide a way to create an abstraction of the sound data that can be searched and stored easily*.

Appendix A

Table of class mapping:

Original	Mapped class
achtergrond verkeer	traffic
achtergrondverkeer	traffic
achterklep	bang
airplane	aircraft
alarm	
ambulance passing	
auto	car
auto afremmen	car
auto die afremt en weer weg rijd	car
auto die ingehaald wordt door moter	car
auto die over een drempel rijd	car
auto die verweg rijd	car
auto drempel	bump
auto heel dichtbij	car
auto in de verte	car
auto met muziek	car
auto met zware motor	car
auto op drempel	bump
auto optrekken	car
auto over hobbel	bump
auto over verkeersdrempel	bump
auto radio	music
auto rijdt langs	car
auto rijdt voorbij	car
auto starten	car
auto ver weg	car
auto's meerdere	car
autoclaxon	horn
autodeaur slaat dicht	bang
autodeur	bang
autoruis op achtergrond	traffic
autos in de verte	traffic
baby	people
bestelbus	car
bicicle	bicycle
bicycle	bicycle
bike	bicycle
bird	birds
birds	birds
birds	birds

Original	Mapped class
bladeren ritselen	rustle
blaffende hond	dog
boem	bang
brakke fiets die voorbij komt	bicycle
brommer	scooter
brommer rijdt voorbij	scooter
bus	bus
busje	car
car	car
car accelerating	car
car door	bang
car horn	horn
car passing	car
cars	car
child	people
claxon	horn
deur	bang
deur dichtslaan	bang
dichtslaande achterbak	bang
dichtslaande deur	bang
dichtslaande deur oid	bang
dierengeluid	
distant alarm horn	horn
door	bang
door opening	
door slammed shut	bang
door slamming	bang
drempel	bump
drempeltje	bump
fiets	bicycle
fiets met klapperende ketting	bicycle
fietser	bicycle
fietser/trappergeluid	bicycle
fietsers	bicycle
fietsketting	bicycle
fietsslot	
fladderen	
fluitend vogeltje	birds
fluitende vogels	birds
fluitje	

Original	Mapped class
footsteps	footsteps
geknerp	rustle
gepraat	people
gerammel	
geritsel	rustle
geritsel van stuikjes	rustle
geschreeuw	people
gesperk	people
gesprek	people
gesprek in de achtergrond	people
getoeter	horn
helikopter	aircraft
hoog fluitend vogeltje	birds
huilende baby?	people
huilende wind	
iemand loopt	footsteps
iemand lopen	footsteps
kind	people
kinderen op de fiets	people
kinderstem	people
kindje roepen	people
knal	bang
kraai	birds
lopen	footsteps
lopen door bladeren	footsteps
lopen in het gras	footsteps
lopen over gras of steen	footsteps
lopen over grint	footsteps
meneer	people
mens op grind	people
mensen	people
mensen lopen	footsteps
mensen praten	people
merel?	birds
mevrouw	people
moter die voorbij rijd	motor
moter in de verte	motor
motor passing	motor
motorcicle	motor
motorcycle	motor

Original	Mapped class
motorvliegtuigje	aircraft
muziek auto	music
nadrukkelijk vogen	birds
opeen volgende autos die over drempel rijden	car
opeenvolgende autos	car
optrekken	car
optrekkende auto	car
optrekkende auto die voorbij komt	car
optrekkende moter	motor
ouoeoe	
passerende auto	car
pedestrian	footsteps
people	people
people talking	people
piepende fiets	bicycle
plane	aircraft
plane landing	aircraft
plane take off	aircraft
poelifinario	
praten	people
praten	people
pratend stel	people
pratende fietsers in de verte	bicycle
pratende mensen	people
pratende mensen op de achtergrond	people
pratende voetgangers	people
ritsel	rustle
roepend kindje	people
roepende man	people
rollerskaters die praten	people
scooter	scooter
scooter die voorbij komt	scooter
snel optrekkende auto	car
snelle auto	car
snelle passerende auto	car
spoor	train
starting car	car
stemmen	people
steps	footsteps

Original	**Mapped class**
talking	people
toeter	horn
toeteren	horn
tok	
tractor	
tractor komt langs	
train passing	train
traplopen	footsteps
trein	train
truck	truck
verkeer	traffic
verkeer in de verte	traffic
vliegtuig	aircraft
vliegtuig/helikopter	aircraft
voetganger	footsteps
voetganger die door grind loopt	footsteps
voetstappen	footsteps
voetstappen op grintpad	footsteps
vogel	birds
vogels	birds
vogels fluiten	birds
vogels in de achtergrond	birds
vogeltjes	birds
vogeltjes tjilpen	birds
voorbij rijdende auto	car
vrachtauto	truck
vrachtwagen	truck
vrachtwagen rijdt langs	truck
walking	footsteps
wandelaar	footsteps
zware auto die voorbij rijd	car
zware wagen denderd	truck

Bibliography

Alain, C. and Arnott, S.: 2000, Selectively attending to auditory objects, *Front. Biosci* **5**, D202–D212.

Andringa, T.: 2002, Continuity preserving signal processing.

Andringa, T.: 2010, Audition: from sound to sounds, *Machine Audition: Principles, Algorithms and Systems* .

Andringa, T. and Niessen, M.: 2006, Real-world sound recognition: A recipe, *Learning the Semantics of Audio Signals* p. 106.

Auvray, M., Gallace, A., Tan, H. and Spence, C.: 2007, Crossmodal change blindness between vision and touch, *Acta psychologica* **126**(2), 79–97.

Bregman, A.: 1990, *Auditory scene analysis: The perceptual organization of sound*, The MIT Press.

Broadbent, D.: 1958, *Perception and communication*, New York: Academic Press.

Brown, G. and Cooke, M.: 1994, Computational auditory scene analysis, *Computer speech and language* **8**(4), 297–336.

Cherry, E.: 1953, Some experiments on the recognition of speech, with one and with two ears, *Journal of the acoustical society of America* **25**(5), 975–979.

Cooke, M.: 1996, Auditory organisation and speech perception: Arguments for an integrated computational theory, *Auditory Basis of Speech Perception*.

Cooke, M. and Ellis, D.: 2001, The auditory organization of speech and other sources in listeners and computational models, *Speech Communication* **35**(3-4), 141–177.

Davies, K.H., B. R. and Balashek, S.: 1952, Automatic speech recognition of spoken digits, *J. Acoust. Soc. Am.* **24(6)**, pp.637 – 642.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J. and Sergent, C.: 2006, Conscious, preconscious, and subliminal processing: a testable taxonomy, *Trends in Cognitive Sciences* **10**(5), 204–211.

Demany, L., Semal, C., Cazalets, J. and Pressnitzer, D.: 2010, Fundamental differences in change detection between vision and audition, *Experimental Brain Research* **203**(2), 261–270.

Driver, J.: 2001, A selective review of selective attention research from the past century, *British Journal of Psychology* **92**(1), 53–78.

Eramudugolla, R., Irvine, D., McAnally, K., Martin, R. and Mattingley, J.: 2005, Directed attention eliminates change deafness in complex auditory scenes, *Current Biology* **15**(12), 1108–1113.

Fellbaum, C., Palmer, M., Dang, H., Delfs, L. and Wolf, S.: 2001, Manual and automatic semantic annotation with WordNet, *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*. !! image annotation !! - ook software tool gebruikt.

Fritz, J., Elhilali, M., David, S. and Shamma, S.: 2007, Auditory attentionfocusing the searchlight on sound, *Current opinion in neurobiology* **17**(4), 437–455.

Gaver, W.: 1993, How do we hear in the world? Explorations in ecological acoustics, *Ecological psychology* **5**(4), 285–313.

Giard, M.-H., Collet, L., Bouchet, P. and Pernier, J.: 1994, Auditory selective attention in the human cochlea, *Brain Research* **633**(1-2), 353 – 356.

Gibson, J.: 1986, *The ecological approach to visual perception*, Lawrence Erlbaum.

Griffiths, T. and Warren, J.: 2004, What is an auditory object?, *Nature Reviews Neuroscience* **5**(11), 887–892.

Grootel, M., Andringa, T. and Krijnders, J.: 2009, DARES-G1: Database of Annotated Real-world Everyday Sounds.

Guastavino, C., Katz, B., Polack, J., Levitin, D. and Dubois, D.: 2005, Ecological validity of soundscape reproduction, *Acta Acustica united with Acustica* **91**(2), 333–341.

Gygi, B. and Shafiro, V.: 2010, Development of the Database for Environmental Sound Research and Application (DESRA): Design, Functionality, and Retrieval Considerations, *EURASIP Journal on Audio, Speech, and Music Processing* **2010**.

Harding, S., Cooke, M. and Konig, P.: 2007, Auditory gist perception: an alternative to attentional selection of auditory streams?, *Lecture Notes in Computer Science* **4840**, 399.

Heller, L. and Skerritt, B.: 2010, Acoustic analysis of perceptual sound categories., *The Journal of the Acoustical Society of America* **127**, 1899.

Hripcsak, G. and Rothschild, A.: 2005, Agreement, the f-measure, and reliability in information retrieval, *Journal of the American Medical Informatics Association* **12**(3), 296–298.

James, W., Burkhardt, F., Bowers, F. and Skrupskelis, I.: 1981, *The principles of psychology*, Harvard Univ Pr.

Kalinli, O. and Narayanan, S.: 2009, Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information, *IEEE transactions on audio, speech, and language processing* **17**(5), 1009.

Klapuri, A.: 2004, Automatic music transcription as we know it today, *Journal of New Music Research* **33**(3), 269–282.

Koch, C. and Tsuchiya, N.: 2007, Attention and consciousness: two distinct brain processes, *Trends in Cognitive Sciences* **11**(1), 16–22. contrasteert met Dehaene 2006.

Krijnders, J.: 2010, Differences between annotating a soundscape live and annotating a recording, Inter-noise 2010, Lisbon, Portugal.

Krijnders, J. and Andringa, T.: 2009a, Soundscape annotation and environmental source recognition experiments in assen (nl), Inter-noise 2009, Ottawa, Canada.

Krijnders, J.D.and Niessen, M. and Andringa, T.: 2009b, Sound event recognition through expectancy-based evaluation of signal-driven hypotheses, *Pattern Recognition Letters* .

Kumar, S., Stephan, K., Warren, J., Friston, K. and Griffiths, T.: 2007, Hierarchical processing of auditory objects in humans, *PLoS Comput Biol* **3**(6), e100.

Levy, M. and Sandler, M.: 2009, Music information retrieval using social tags and audio, *Multimedia, IEEE Transactions on* **11**(3), 383 –395.

Lin, C., Naphade, M., Natsev, A., Neti, C., Smith, J., Tseng, B., Nock, H. and Adams, W.: 2003, User-trainable video annotation using multimodal cues, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM, pp. 403–404.

Mack, A.: 2003, Inattentional blindness, *Current Directions in Psychological Science* **12**(5), 180.

Mack, A. and Rock, I.: 1998, *Innatentional Blindness*, Cambridge, MA: MIT Press.

McAnally, K., Martin, R., Eramudugolla, R., Stuart, G., Irvine, D. and Mattingley, J.: 2010, A Dual-Process Account of Auditory Change Detection, *Journal of Experimental Psychology: Human Perception and Performance* **36**(4), 994–1004.

Nelken, I. and Ahissar, M.: 2006, High-level and low-level processing in the auditory system: the role of primary auditory cortex, *Dynamics of speech production and perception* p. 343.

Oliva, A.: 2005, Gist of the scene, *Neurobiology of attention* **17**.

Pavani, F. and Turatto, M.: 2008, Change perception in complex auditory scenes, *Perception and Psychophysics* **70**(4), 619.

Rensink, R., O'Regan, J. and Clark, J.: 1997, To see or not to see: The need for attention to perceive changes in scenes, *Psychological Science* **8**(5), 368.

Schafer, R.: 1977, *The tuning of the world*, Alfred A. Knopf.

Scott, S.: 2005, Auditory processing–speech, space and auditory objects, *Current Opinion in Neurobiology* **15**(2), 197–201.

Shamma, S.: 2001, On the role of space and time in auditory processing, *Trends in Cognitive Sciences* **5**(8), 340–348.

Shinn-Cunningham, B.: 2008, Object-based auditory and visual attention, *Trends in cognitive sciences* **12**(5), 182–186.

Simons, D. and Chabris, C.: 1999, Gorillas in our midst: Sustained inattentional blindness for dynamic events, *PERCEPTION-LONDON-* **28**, 1059–1074.

Simons, D. and Rensink, R.: 2005, Change blindness: Past, present, and future, *Trends in Cognitive Sciences* **9**(1), 16–20.

Sinnett, S., Costa, A. and Soto-Faraco, S.: 2006, Manipulating inattentional blindness within and across sensory modalities, *The quarterly journal of experimental psychology* **59**(8), 1425–1442.

Tzanetakis, G. and Cook, P.: 2000a, Audio information retrieval (AIR) tools, *International Symposium on Music Information Retrieval*.

Tzanetakis, G. and Cook, P.: 2000b, Experiments in computer-assisted annotation of audio, *International Conference on Auditory Display (ICAD), Atlanta, Georgia, USA*.

Tzanetakis, G. and Cook, P.: 2000c, Marsyas: A framework for audio analysis, *Organised sound* **4**(03), 169–175.

Van Hengel, P. and Andringa, T.: 2007, Verbal aggression detection in complex social environments, *IEEE Conference on Advanced Video and Signal Based Surveillance, 2007. AVSS 2007*, pp. 15–20.

Vitevitch, M.: 2003, Change deafness: The inability to detect changes between two voices, *Journal of Experimental Psychology, Human Perception and Performance* **29**(2), 333–342.

Wang, D. and Brown, G.: 2006, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H.: 2006, Elan: a professional framework for multimodality research, *Proceedings of Language Resources and Evaluation Conference (LREC)*, Citeseer.

Young, S.: 1996, Large vocabulary continuous speech recognition: A review, *IEEE Signal Processing Magazine* **13**(5), 45–57.

Zajdel, W., Krijnders, J., Andringa, T. and Gavrila, D.: 2007, CASSANDRA: audio-video sensor fusion for aggression detection, pp. 200–205.