Signal-driven sound processing for uncontrolled environments

Dirkjan Krijnders

Front cover: ...
Back cover: ...
Printed by: ...
ISBN: ...

university of groningen

NWO
Nederlandse Organisatie voor Wetenschappelijk Onderzoek

bcn Graduate School for Behavioral and Cognitive Neurosciences

This research was made possible thanks to NWO grant 634.000.432 (To-KeN/CASSANDRA).

Rijksuniversiteit Groningen

Signal-driven sound processing for uncontrolled environments

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
dag **datum 2010
om **tijd** uur

door

Johannnes Dirk Krijnders

geboren op 2 mei 1978
te Amsterdam

Promotor:	Prof. dr. L.R.B. Schomaker
Copromotor:	Dr. T.C. Andringa
Beoordelingscommissie:	Prof. R. K. Moore, PhD Prof. Dr.-ing. H. Fastl Prof. Dr. Ir. D. Botteldoorn

Contents

I Introduction **1**

1 Introduction **3**
1.1 Application domains 4
1.1.1 Ambient awareness 5
1.1.2 Soundscapes 6
1.2 Automatic speech recognition as perceptive system 6
1.3 Perception 7
1.4 Research questions 8

2 Basic signal processing **11**
2.1 Design criteria 11
2.1.1 Sound sources 11
2.1.2 Concurrent sound sources 12
2.1.3 Transmission effects 13
2.1.4 Unknown sources 15
2.2 On tones and pulses 15
2.3 Cochleogram 16
2.3.1 Basilar membrane 16
2.3.2 Cochleogram 18
2.4 Local target to non-target ratio 19
2.5 Scope of this thesis 19

II Theory of sound source recognition in uncontrolled environments 23

3 Tone, pulse, and chirp decomposition 25
3.1 Background . . . 28
3.1.1 ASA and Auditory Objects . . . 28
3.1.2 The local signal-to-noise ratio . . . 28
3.1.3 Required properties . . . 30
3.2 Methods . . . 30
3.2.1 Time-frequency processing . . . 31
3.2.2 Standard deviation of noise . . . 31
3.2.3 Tone and pulse fit . . . 32
3.2.4 Separating noisy fluctuations from “real” signal contributions 32
3.3 Experiments . . . 34
3.3.1 A comparison with energy threshold . . . 35
3.3.2 Tone-fit and pulse-fit for chirps . . . 35
3.3.3 Correlation with local signal-to-noise ratio . . . 38
3.3.4 Tone sensitivity . . . 39
3.3.5 Spectral and temporal accuracy . . . 39
3.3.6 Proximate and crossing tones . . . 39
3.3.7 Proximate pulses . . . 43
3.3.8 Recorded sound sources . . . 44
3.4 Discussion . . . 49
3.4.1 Broadband residue . . . 49
3.4.2 Relation with sparse modeling techniques . . . 50
3.4.3 Relation with human processing . . . 51
3.5 Conclusion . . . 51

4 Sound event identification through expectancy-based evaluation of signal-driven hypotheses 53
4.1 Signal-driven processing . . . 56
4.1.1 Dataset . . . 56
4.1.2 Signal components . . . 56
4.1.3 Harmonic complexes . . . 58
4.1.4 Broadband events . . . 58
4.2 Dynamic network model . . . 59
4.2.1 Knowledge network . . . 61
4.2.2 Dynamic network of hypotheses . . . 62
4.2.3 Activation . . . 63
4.3 Experiments . . . 64
4.3.1 Experimental setup . . . 65
4.3.2 Signal-driven results . . . 66
4.3.3 Expectancy-based results . . . 67
4.4 Discussion . . . 67

5 How to evaluate the sources in a soundscape? 69
5.1 introduction . . . 69
5.2 Methods . . . 71
5.2.1 Dataset . . . 72
5.2.2 Preprocessing . . . 72
5.2.3 Feature vectors . . . 74
5.3 Results and discussion . . . 76

III Evaluation 79

6 Ambient awareness: Aggression detection 81
6.1 System description . . . 83
6.1.1 Audio unit . . . 83
6.1.2 Video unit . . . 84
6.1.3 Fusion unit . . . 86
6.2 Experiments . . . 89
6.3 Conclusions and future work . . . 92

7 Evaluation on a dataset of city sounds 95
7.1 Methods . . . 96
7.1.1 Sound Processing . . . 96
7.1.2 Dynamic Network Model . . . 96
7.2 Experiment . . . 96
7.2.1 Data . . . 97
7.2.2 Setup . . . 97
7.2.3 Results . . . 98
7.3 Conclusions . . . 99

8 Automatic Extraction of Formants in Noise 103
8.1 Introduction . . . 103
8.2 Method . . . 105
8.2.1 Algorithm . . . 105
8.2.2 Material . . . 105
8.2.3 Evaluation . . . 105
8.3 Results . . . 107
8.3.1 Formant extraction . . . 107
8.3.2 Vowel classification . . . 108
8.4 Discussion . . . 109
8.4.1 Harmonic complex extraction . . . 109
8.4.2 Formant extraction . . . 110
8.4.3 Vowel classification . . . 110
8.5 Conclusion . . . 111

9 Conclusions **113**
9.1 Future work 116
9.1.1 Signal Processing 116
9.1.2 Annotation 117
9.1.3 Recognition 118

10 Samenvatting **119**

A Gamma-chirp in formulae **121**

B List of publications **123**
B.1 Journal papers 123
B.2 Conference proceedings 123

Acknowledgements **125**

References **127**

Part I

Introduction

1

Introduction

The human auditory system plays an important role in the ambient awareness of people. It warns you against danger when a car comes from behind and relaxes if hear just birds singing and leaves rustling. The ease with which humans perform these task is in strong contrast with the problems of automatic sound recognition systems. However, these automatic sound recognition systems can play an important role in our future lives through a range of applications. From improved human-computer interfaces and assisting the deaf, to automatically estimating the sound quality of quiet areas. Sound source recognition forms the basis of these technologies. For example, a computer may stop responding to voice commands when it detects you are using your phone. The watch of a hearing impaired person may warn him that a car is approaching from behind. And even though the average sound level in a quiet area may be low, if some highly disturbing sounds occur a few times per day the "quiet" area may not be so nice at all.

However, current technology is limiting sound recognition to voice-response systems and awkward-to-use dictation software. The reason for these limitations are the basic assumptions that are necessary to ensure that the technology works at acceptable performance levels. However, when these systems are moved from highly controlled environments to more realistic environments these assumptions become constraints. There are four main constraints. The first is that the current systems assume that the incoming signal stems from a single source, while in realistic environments any number of sources can co-occur. As a result playing music in the background while dictating confuses the recognition system because it assumes that the

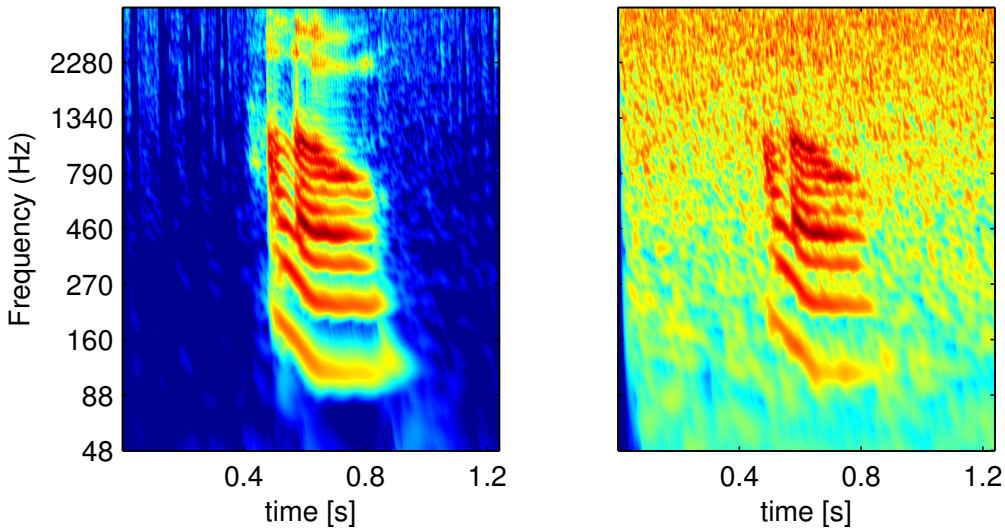


Figure 1.1: On the left a time-frequency representation (see section 2.3) of the utterance "hallo" by the author in clean conditions. On the right the same utterance with added noise that is as energetic as the speech. The sample on the right is often used in speech recognizer tests.

music is part of your voice. The second constraint is the assumption that the system can recognize the incoming signal, while in realistic environments unknown sound classes are likely to occur. In the case of the music playing in the background, this would mean that the software tries to recognize the music as speech. The third constraint is that the acoustic conditions around the microphone are constant and known. This entails that after the dictation software is trained the microphone must not be changed or moved relative to the source. This is why headsets are commonly sold with dictation software. The last constraint is that noise sources, either in the system itself or outside the system, like air-conditioning, must have certain "nice" properties which allow the system to discount them. The right image in figure 1.1 shows an example of an utterance in "nice" noise, figure 1.2 shows the same utterance in a more realistic background.

The rest of this chapter is devoted to applications and why current techniques are too limited to solve the problems posed by these applications. The chapter concludes with the research questions addressed in this thesis.

1.1 Application domains

Sound source recognition has many potential applications. Three main applications on which the evaluation part of this is thesis is focused are ambient awareness issues, like aggression detection and noise control, and soundscape research.

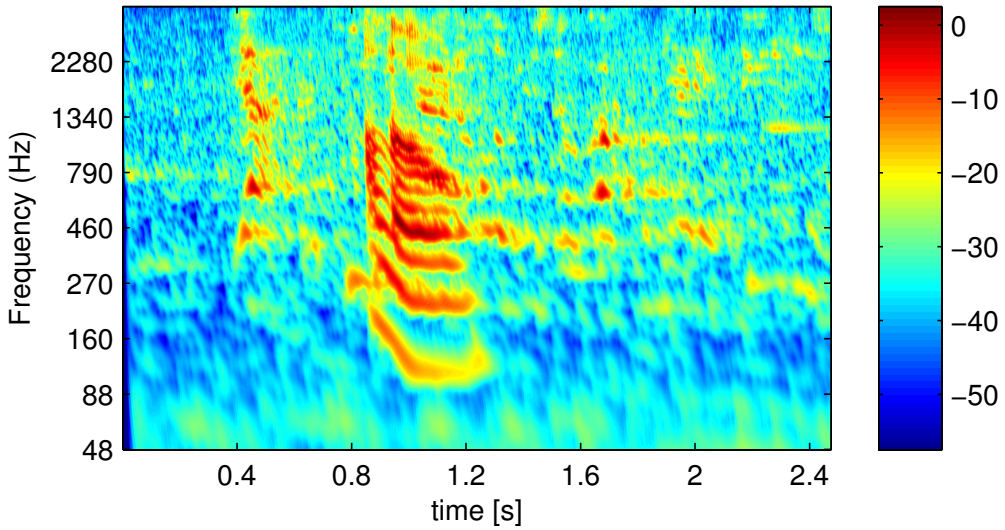


Figure 1.2: The same speech as in figure 1.1, but now mixed in a realistic sound signal recorded on the Amstel Station in Amsterdam.

1.1.1 Ambient awareness

Many city centers are watched via closed circuit video cameras and these camera feeds can both be used for monitoring as well as forensic research after a crime. Monitoring camera feeds is a very intensive task and often a single observer is responsible for many feeds. These factors increase the risk of missing a relevant event. One of the main motivations for monitoring is public safety in entertainment districts and the main target is preventing aggression. As aggression is often preceded by aggressive screams, sound identification can help to prioritize camera feeds for the operator. These kind of systems are already installed in some cities and perform acceptable (van Hengel and Andringa, 2007).

Another example of ambient awareness is the monitoring of noise control laws. This monitoring is currently purely based on the basis of level measurements. These sound levels, expressed in a single number, like $L_{A,eq}$, are easy to measure and to compare with regulations of these levels. However a level measurement does not tell anything about who or what made the noise. Recognizing what caused the sound will improve the usefulness of the monitoring.

Both applications require that (hopefully) rare events are detected, as the goal is to reduce these events. Detecting these rare events is a challenge as 99.9% of the time the events are to be ignored (van Hengel and Andringa (2007) reports 2 events (aggressive screams) in 18 days, i.e. 30 seconds in 26000 seconds). In general reducing the number of missed events increases the number of false alarms as this usually entails more permissive settings.

1.1.2 Soundscapes

When people are asked to describe a visual landscape their answers are often a general description of the scene, sometimes with a few (salient) details mentioned as well. For example one may say, this is a flat country side with cattle and there is a red car parked at that farm. For the sonic environment a similar question can be posed and the answers are a similar mix between general description and the odd sounds. Such a perceived sonic environment is called a soundscape (Schafer, 1977).

Soundscape research has focussed on describing the way people perceive the soundscape and how the soundscape influences human behavior. Soundscapes can influence people both in a positive (Irvine et al., 2009) way, such as in parks and churches and a negative way (Truax, 2001, Chapter 6), such as in cities with traffic and construction work. Models of this influence exist mainly for the negative case (De Coensel and Botteldooren, 2008), but these do not include the essential step of sound source recognition. However sound source recognition is an essential basis for these models (De Coensel and Botteldooren, 2008; COST Action 0804, 2008, page 17), both to generate a high-level summary of the soundscape as well as identify the individual sources.

1.2 Automatic speech recognition as perceptive system

To illustrate the constraints mentioned in the introduction, the next example shows what would happen if our auditory system functioned like current automatic (speech) recognizer systems. The reasons for the weird behavior follow after the example:

Suppose you're standing at a station platform and a rather nice girl (or boy/woman/man) walks up close to you and starts talking to you, although you speak the same language she has to start with at least ten standard sentences before you can actually understand anything else she'll say to you. Of course you have to speak the same ten sentences before she can understand you. After these formalities you actually start a conversation, until a train enters on the other platform. What used to be fairly comprehensive sentences with few recognition errors, changes into complete jabberwocky for both. Even when the noise of the train has subsided, you'll have repeat your ten sentences in order to understand each other again.

People probably would not be using speech if the above example was taken from our daily lives. So why does an automatic speech recognizer work so differently? The girl has to speak at close range from you to minimize the influence of the acoustics of the train station and even then you need to adjust your automatic speech recognizer to the acoustic environment and

her voice specifics. The noise of the approaching train mixes with the speech and the speech recognizer tries to recognize the combination, which results in the jabberwocky. In addition the train changes the acoustics of the station so both of you need to readjust to the new acoustical environment.

So the basic constraints in a modern speech recognizer are, first, the assumption that the input signal stems from a single source, possibly mixed with a type of noise that the recognizer is trained to ignore, so that if you add train noise the recognition results will drop. Second, the assumption that the acoustical environment is known, hence the ten sentences to learn the new acoustic environment, and stable, hence the requirement to relearn the acoustical environment after the train enters.

For the current applications of speech recognizers, such as voice response systems and dictation software, these limitations are not a constraint, because the system implicitly or explicitly forces the user to make sure these limitations are met. However if we want to move from these application to less constrained environments and from only speech to all possible sound classes around us, these limitations are show-stoppers.

1.3 Perception

In contrast to the situation described in the previous section the human auditory system has no problems to function in uncontrolled environments. Some of the problems mentioned above are not even noticed by human listeners. For example, humans do not hear the acoustics of a room, unless they are trained to do so or the acoustics are extreme (Nábělek and Robinson, 1982). Nor do they have much trouble in situation like a cocktail party, where many voices are present but conversation is still possible (Bronkhorst, 2000; Cherry, 1953). Insight in how the human auditory system functions may improve our methods in automatic source recognition.

The human auditory system seems to group energy belonging to a single source. For example, if we hear a car, we hear just that. Only when we make an effort we can separate the sound from the tyres from the sound of the engine (Gaver, 1993a; Shinn-Cunningham, 2008). If sound from a single source arrives at the ear in isolation this grouping is trivial, but when other sources are present the grouping becomes harder, but also more essential, because if we have grouped the energy there is a pattern of energy to identify the source instead of separate parts. On the other hand, grouping in the presence of multiple sources becomes easier if we know which sources are present. This is a paradox because we need to know the sources present to group correctly, but we need the groups to recognize the sources correctly. The paradox is called the signal-in-noise paradox (Andringa, 2002).

Psycho-acousticians have researched how this grouping occurs at early stages of processing (Bregman, 1990). Although this field is called “auditory

scene analysis", it has mainly concerned itself with artificial tones and how they are grouped in harmonic complexes (or complex tones) or under which conditions the auditory system perceives tones as a single stream or as two or more streams when played in succession at different rates and frequencies (van Noorden, 1975). For example in music the bass-line and melody are conceptual "streamed" in different perceptual streams, even though they are played on the same instrument.

However research on how this grouping works in more realistic settings is a recent development. In these settings grouping is not limited to complex tones but, for example, extends to the sound of a car as a single percept. This last grouping is analogous to the formation of objects in the visual domain and thus is called a auditory object. However, a definition of an auditory object is difficult (Shinn-Cunningham, 2008), in part because of what constitutes an object in human perception is strongly dependent on the state of listener. However Shinn-Cunningham (2008) gives the following working definition of a auditory object: An auditory object *"is an estimate of sound emanating from a discrete sound source: an 'auditory object' is a perceptual entity that, correctly or not, is perceived as coming from one physical source."*.

Griffiths and Warren (2004) agree with this definition, but warn that equating an auditory object to a sound source is arbitrary. It may as well be a sound event. For example, the utterance "a" may form an object of the voice, which relates to the source, and an object of the vowel /a/ which is more related to its role in speech. Which of these classes of objects the system forms dependents on the task of the system. Note also that auditory objects can be formed at many levels of abstraction. In this thesis most formed are close to the signal, such that they can be formed using signal-driven heuristics.

If we can form hypotheses of auditory objects they can form the basis of solving the signal-in-noise-paradox by hypothesizing how to group. These hypotheses can be formed in a signal-driven way, i.e. what we extract from the signal creates hypotheses. This is the subject of this thesis. The hypotheses could also be formed in a knowledge driven way, for example after a train whistle a departing train is expected. This is the subject of the PhD thesis by M.E. Niessen (Niessen, 2010). The final system should integrate both ways of generating hypotheses.

1.4 Research questions

To build to systems mentioned in section 1.1 we need techniques that do not make the assumptions or constraints mentioned in the beginning of this chapter. To develop these techniques we need to answer the following research questions:

1. *How to select sonic evidence that is highly likely to stem from a single source from a sound signal recorded in realistic acoustic circumstances?*
2. *How can the input signal, instead of the system design, guide the processing of the signal, towards an optimal rendering of the information in the signal??*

Answering the first question should alleviate the assumption that there is only one source present and that we can do this in arbitrary acoustical circumstances. This is in line with the perceptual formation of auditory objects.

The answer to the second question frees the system from choices made during the design of the system. This should alleviate the assumption that the input can be recognized: if part of the input is detected to be music, the system should not apply knowledge of speech to recognize that part of the input.

The next chapter will introduce the signal processing as used throughout this thesis. The chapter concludes with a further overview of this thesis.

2

Basic signal processing

2.1 Design criteria

Most applications mentioned in section 1.1 require a system that must work out-of-the-box, i.e. without the intervention of a specialist. The system should work in cities, villages, countryside and should be undisturbed by influences of the weather or other changes in the environment. For most applications it is not necessary to recognize all sound sources perfectly, but the general pattern of sound sources should provide a reliable indication of the sonic environment. These requirements pose demands on the design of the system:
The system should be able to

- handle concurrent sources (section 2.1.2)
- deal with changing transmission effects (section 2.1.3)
- handle diverse, possible unknown, classes (section 2.1.4)

The next section will discuss what a (environmental) sound source is and the following sections describe the problems associated with the demands mentioned above.

2.1.1 Sound sources

Based on the dictionary (New Oxford American Dictionary) definitions of "sound" and "source" a sound source would be "a place, person, or thing

from which something comes", where something is "vibrations that travel through the air or another medium and can be heard when they reach a person's or animal's ear". This would entail that a sound source is the place where the vibrations originate, but this "place" is unclear when considering the common sense definition that for example a car is a sound source, while the sound source would technically be the explosions in the engine, the air flow around the car and the noise from the interaction between the tires and the road. The sound source is thus, like the auditory objects in section 1.3, dependent on the detail level the user is interested in.

In some literature a further qualifications like "environmental" (Cowling and Sitte, 2003; Shafiro and Gygi, 2004), or "everyday" are given to a sound source. But clear definitions of these qualifications are lacking. It is usually assumed that environmental sounds exclude speech or music, however these terms are not clearly defined either. Ballas and Howard (1987) and Vanderveer (1979) agree on two criteria for environmental sounds: First they are "produced by real events" and second they convey "meaning by virtue of the causal events".

2.1.2 Concurrent sound sources

As sound sources seldom occur in isolation in the real world it is important to handle concurrent sources. Usually this problem is treated as a sound source with added noise (Hayes, 1996, Chapter 7):

$$x(t) = x_{target}(t) + n(t) \tag{2.1}$$
$$n(t) \sim \mathcal{N}(\mu, \sigma^2) \tag{2.2}$$

This approach assumes that the non-target sources add up to normally distributed noise, which is only true for certain (noisy) sources or when many uncorrelated sources are present (central limit theorem). The assumption can be valid in cases where there is much control over the recording and the main noise source is known, for example a telephone with close talking microphone increases the probability of a single source and the noise may be caused by electrical and thermal sources which can be assumed to be broadband noise. The more general problem can be stated as (Cardoso and Martin, 2007):

$$x(t) = \sum_{n=1}^{N} a_n x_{s,n}(t) \tag{2.3}$$

where N is the number of sound sources, a_n represents the sound level decrease due to the distance between the sound source and the receiver and $x_{s,n}(t)$ is the time signal of an individual source at that source. This problem is a standard inverse problem and cannot be solved without extra knowledge

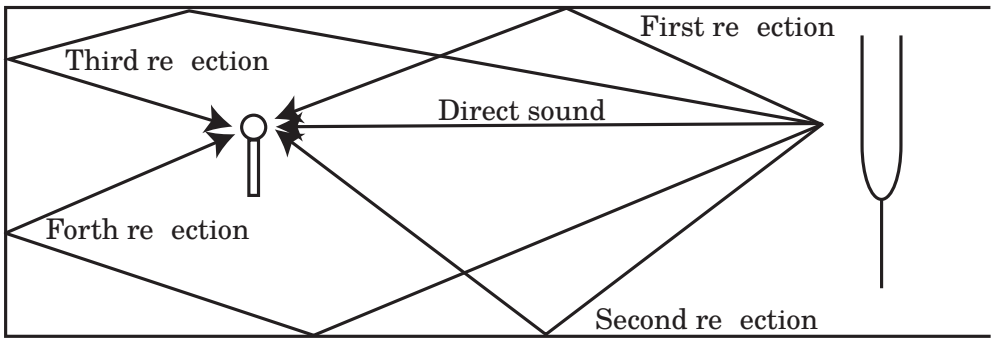


Figure 2.1: The path of the direct sound and the first four reflections in a room.

if the number of microphones (observations of $x(t)$) is less then the number of sources (N), i.e. the system is underdetermined. Hence, most current approaches use multiple microphones and independent component analysis (ICA, (Choi et al., 2005)) or beam-forming techniques(Kellerman, 2009) to extract to separate $x_n(t)$. This changes equation 2.3 to

$$x_m(t) = \sum_{n=1}^{N} a_{n,m} x_{s,n}(t) \tag{2.4}$$

$$\mathbf{x}_m = A\mathbf{x}_s \tag{2.5}$$

where $A = a_{n,m}$ and the matrix formulation is the standard problem statement of ICA. And this problem is solvable when the matrix A is constant (or slowly changing) and only one of the sound sources $\mathbf{x}_s$ produces broadband noise. This entails that the range of application is limited to locations where the acoustics are (quasi-)constant and the sources are (quasi-)stationary. This prevents ICA from being applicable in the applications mentioned in section 1.1.

2.1.3 Transmission effects

As sound travels from the source to the microphone it is not only attenuated and mixed with other sources, the sound also reflects off surfaces leading the multiple paths from source to microphone 2.1. The indirect paths are longer and thus a delayed and more attenuated version of the original sound arrives at the microphone. Apart from being delayed, the frequency content of the sound changes as surfaces don't necessarily reflect all frequencies equally well.

Reverberation is usually modeled as an impulse response. It can be obtained by recording of the sound of a perfect impact or a swept sine (Farina, 2005), or be derived by modeling the room and the object inside it. The sound arriving at the microphone can be simulated by convolving the source sound

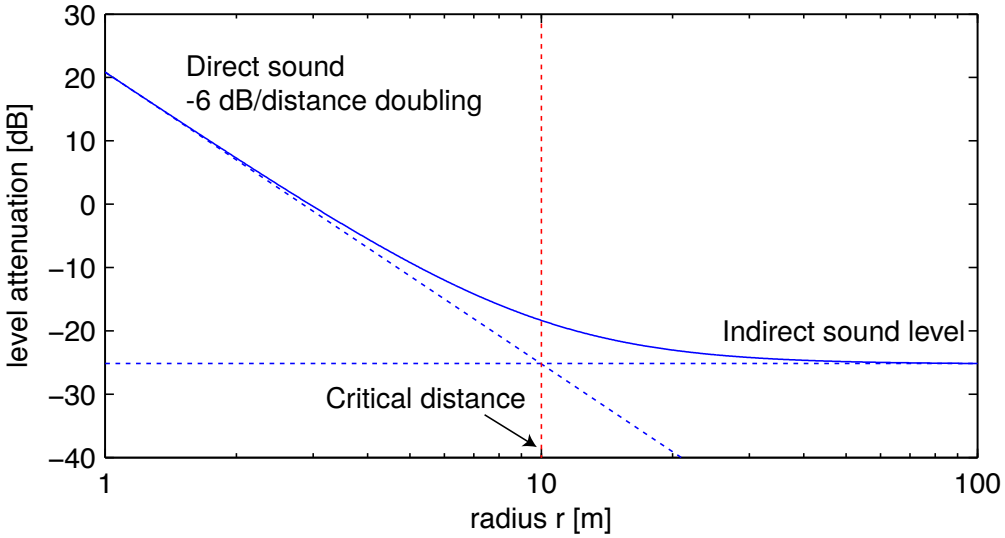


Figure 2.2: The attenuation of the sound pressure level in a ideal room of $50m^3$. The critical distance is the distance where the level of the indirect sound is equal to the direct sound, in this case $10m$.

with the impulse response, thus replacing the simple factor in equation 2.3 with a convolution:

$$x_m(t) \quad = \quad \mathbf{r} \otimes x_s(t) \tag{2.6}$$

The impulse (r) is dependent on both the position of the source and the receiver and on the properties of the room and objects in it. If source and receiver are close in comparison to the size of the room the direct sound contributes most energy to the receiver. As source and receiver move away from each other the relative contribution of the direct sound compared to the indirect sound (constant) decreases (see figure 2.2). The distance from the source at which the contribution for the indirect sound equals the contribution from the direct sound is called the critical distance. Within this ratio speech is understandable independent of the amount of reverberation, outside the intelligibility is a function of the amount of reverberation (Peutz, 1971). Our system will need to work both inside (Chapter 8) and outside (Chapter 6 and 7) the reverberation radius.

Reverberation complicates the mixing of sources by adding time-delayed, frequency-dependently attenuated copies each source. The only interval that the sound is undisturbed by delayed copies is when the direct sound has arrived but first reflection has not. These intervals are very short for rooms (3 ms = 1 meter path length difference). Yet this property is exploited by humans for source localization and source identification. This exploitation is called the precedence effect (Litovsky et al., 1999). To exploit this effect the

system needs to detect the unperturbed start of the sound, current systems are not build, nor suitable, to do that.

Current techniques to deal with reverberation can be split in two categories (Habets, 2007): reverberation suppression and reverberation cancelation. This last category requires a full estimate of the impulse response and is therefore unsuitable for changing environments (Haykin, 1994, 2000). Reverberation suppression requires less knowledge of the transmission properties, but needs more information about the sources present as these methods use knowledge of the source. This knowledge can include the kind of source (Deller Jr. et al., 1999), e.g. speech for linear prediction coding, or location for beam-forming techniques (Trees, 2002). Since the sources in an open environment are unknown, this requirement is a problem.

2.1.4 Unknown sources

When the system is deployed in uncontrolled environment it will encounter sound sources that it has no knowledge of. It will not be able to classify these sources, but it would be beneficial if the recordings of these sources are stored for further analysis. Current systems expect that all encountered classes were part of their training database. This entails that they assign a class to every part of the signal they try to classify, regardless of whether that class actually matches the signal, but it just happened to have a slightly higher probability then the other classes. The only mechanism those systems have to ignore unknown sources is to ignore sources that do not exceed a certain threshold of probability.

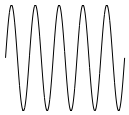
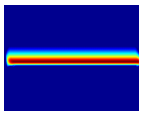
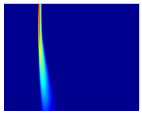
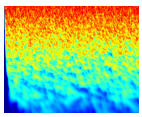
2.2 On tones and pulses

To extract evidence that is likely to stem from a single source, we need a criterium that allows us to do so, solely based on the signal at hand. Sound production can roughly be divided into three different mechanisms: resonance, impact and turbulence (Gaver, 1993b). These mechanisms result in respectively tones, pulses and broadband noise[1]. The first two are very localized, tones in frequency and pulses in time. This makes that the chance for two tones or pulses to mask each other in the time-frequency plane is small. As such tones and pulses that stem from a single source have a large chance to be found as a single time-frequency region.

These three mechanisms match the extremes of the Heisenberg inequality

[1] The term “noise” will be used in this thesis in the meaning of colored or white noise, i.e. noise resulting from aperiodic processes. Compare to “ruis” in dutch or “Krach” in german (Dubois and Guastavino, 2008)

Table 2.1: Limits of the Heisenberg inequality

Tones	$\lim_{\sigma_t \to 0} \sigma_t \sigma_\omega = \frac{1}{2}$		
Pulse	$\lim_{\sigma_\omega \to 0} \sigma_t \sigma_\omega = \frac{1}{2}$		
Broadband	$\lim_{\sigma_t \sigma_\omega \to \infty}$		

(see table 2.1). The Heisenberg inequality states that:

$$\sigma_t \sigma_\omega \quad \geq \quad \frac{1}{2} \tag{2.7}$$

where σ_t is the time variation and σ_ω is the frequency variation. This notation follows Hut et al. (2006), but the inequality was first derived by Gabor (1946). Gröchenig (2001, Chapter 2) discusses the uncertainty relation extensively.

The localization of tones and pulses allows for tracking them through time, resp. frequency. This tracking provides extra certainty and groups together similar parts of the spectrum.

2.3 Cochleogram

As we are interested in tones and pulses and we want to track them we need a continuous development of tones and pulses through our time-frequency representation without noticeable biases for special frequencies or points in time. Time-frequency representations decompose the audio signal into two-dimensional matrices where time is one dimension and frequency the other. The most popular one is the spectrogram based on the short-term fast Fourier transform (SFFT). However this Fourier transform does have biases, both in time and frequency, see figure 3.13 . An alternative is to use a basilar membrane or cochlea model which does not have these biases and thus makes it easier to track continuous signals through time and frequency.

2.3.1 Basilar membrane

The basilar membrane is the last mechanical part of the human auditory system. Its movement triggers responses in the haircells on its surface. The

Figure 2.3: Location of the cochlea (left, purple) and a cross-section of the cochlea (right) along the gray line. The location of the basilar membrane is highlighted in red . The membrane vibrates when sound reaches the ear and the haircells on the membrane translate the amplitude change to nerve signals.

basilar membrane divides the cochlear duct in two halves (figure 2.3). This cochlear duct is rolled in a snail-like shape end reduces in diameter from the entrance to the end. This reduction changes the frequency properties of the basilar membrane and thus different frequencies in the sound stimulate different haircells.

Many models of this system have been proposed, either with transmission-line models which model the physics of the cochlea (Duifhuis et al., 1985) or with filterbank models (Irino and Patterson, 1997) which try to match the psycho-acoustical data.

The transmission-line model has shown to be more accurate in resolving sound close to the Heisenberg limit than the gamma-tone filterbank Hut et al. (2006). So for our purposes the transmission-line model would be better, however due to the wide-spread use of the gamma-tone filterbank we use the gamma-tone filterbank through out this thesis. Most experiments have also been done with the Duifhuis et al. (1985) model and the results were very similar.

The gamma-tone or gamma-chirp filterbank (Irino and Patterson, 1997) is a widely used model for the basilar membrane. Its filter coefficients (h_{gc}) are defined by ($c = 0$ for the gamma-tone):

$$h_{gc} = at^{N-1}e^{-2\pi bB(f_c)t}e^{j(2\pi f_c t + c\log(t))} \tag{2.8}$$

where f_c is the center frequency of the channel, N the order of the gammatone (4) and $a = 1$, $b = 0.71$ and $c = -3.7$. These values are somewhat different than in Irino and Patterson (1997) in favor of a narrow tonal response (at the cost of increased group delay) to make its response closer to

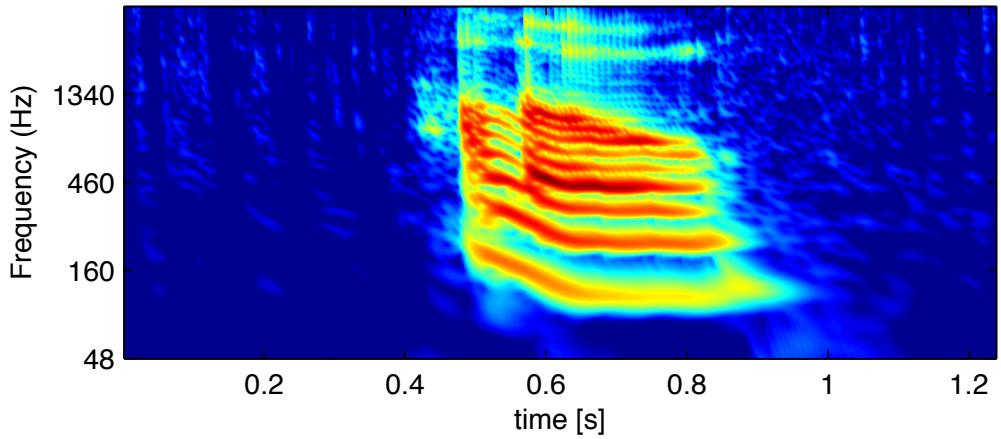


Figure 2.4: Cochleogram of the author saying "hallo", the noise is from surroundings

that of the transmission-line model. The frequency range extends from 60 Hz to 4000 Hz. The center frequencies of the filterbank are distributed logarithmically. The bandwidth of each filter is given by the ERB scale (Moore and Glasberg, 1996):

$$B(f_c) \quad = \quad 24.7 + 0.108 f_c \tag{2.9}$$

2.3.2 Cochleogram

The basilar membrane models result in a amplitude representation of the membrane, this gives very detailed information on the signal which may be useful (Krijnders et al., 2007). However for our purposes a energy representation is more convenient. To calculate the energy in the amplitude matrix the filter output is squared and leaky integrated with a channel dependent time-constant $\tau_c = \max(5, 2/f_c)$ ms. This leaky-integration method yields, in combination with the logarithmic frequency axis, a constant-Q-like representation . The filterbank output is squared to represent an energy measure, down-sampled to 200 Hz, and compressed logarithmically to express the energy in dB. The resulting representation is a spectrogram-like representation, termed a cochleogram, with 5 ms frames. For a full mathematical description see appendix A.

Both the frequency and the energy representation are logarithmic and comply with Weber's law, which entails that they are able to represent many orders of magnitude in a limited dynamic range. Both are central properties of auditory processing. For historical reasons "channels" are often referred to as "segments" and we will use both terms.

2.4 Local target to non-target ratio

In speech recognition research the signal to noise ratio (SNR) is often used as a measure of how bad a signal is degraded by mixing with other sources (noise) than the target signal. It is defined as the ratio of the power of the signal to the power of the noise:

$$\text{SNR} = 10 \log \left(\frac{P_{signal}}{P_{disturbance}} \right) \tag{2.10}$$

This ratio is calculated for a complete speech sample and thus a global measure. However spectral and temporal content of the signal and the noise are often different, so the SNR says nothing about the ratio of power at specific moments in time and frequency. This problem is sometimes alleviated by using A-weighting (S1.4, 2001, based on Fletcher and Munson, 1933) both the speech and noise signal and using speech detection (P.56, 1993). This last method prevents the silences in speech from changing the SNR values.

From speech research started by Fletcher (Fletcher, 1950) and continued by others (Allen, 1994; Bronkhorst, 2000) it is known that only the local target to non-target ratio counts in human performance. Cooke (2006) proposed a model for speech recognition that uses this fact under the term glimpsing at the target when the LSNR ratio is advantageous to do so.

Moreover, and this is more a naming question, it assumes that signal and noise are well defined, but if two people talk at the same time both may be considered signal. For this reason the term target-to-non-target ratio is more appropriate.

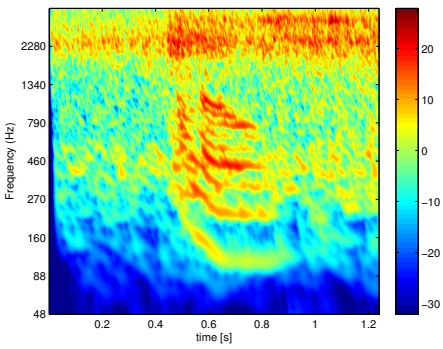
Figures 2.5(e-f) show the LSNR of the signal “hallo” with pink, respectively white, noise added at 0 dB target to non target ratio. Although the signals in figures 2.5(c-d) don’t include any weighting or speech activity detection, the net effect is a slight overestimation of the SNR compared to figure 2.5(a).

$$\text{LSNR}(t, f) = 10 \log \left(\frac{P_t(t, f)}{\sum_{n \neq t} P_n(t, f)} \right) \tag{2.11}$$

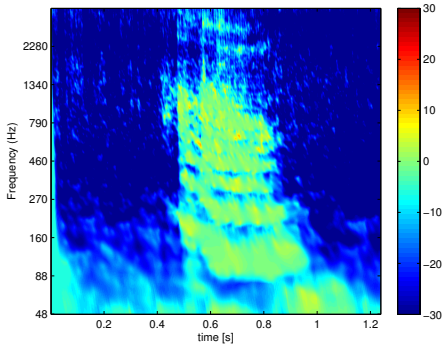
where P_t is the energy of the target sound and P_n is the power of all individual sources.

2.5 Scope of this thesis

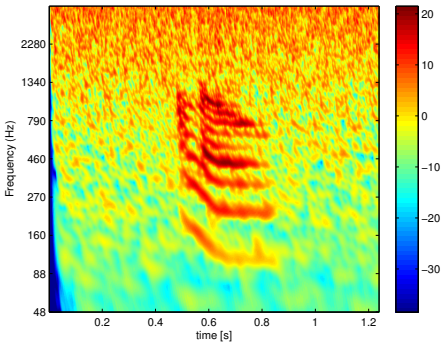
Based the design criteria (section 2.1) we formulate the goals of the work presented in this thesis.



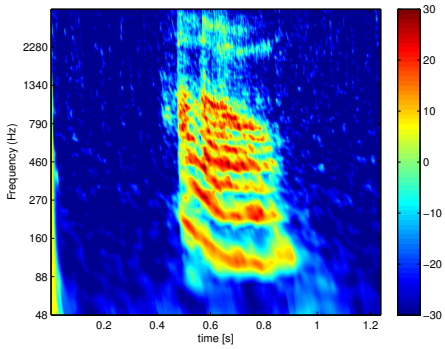
(a) Signal from figure 2.4 with subway noise added at zero dB SNR following A-weighting and ITU-T P.56 speech detection



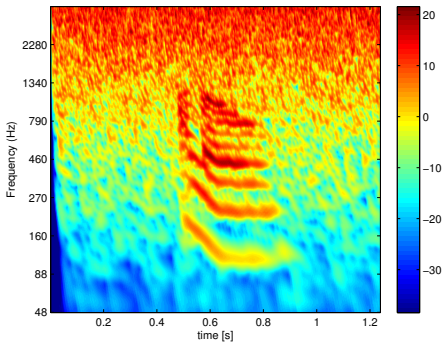
(b) Local target to non target ratio of the signal in figure 2.5(a)



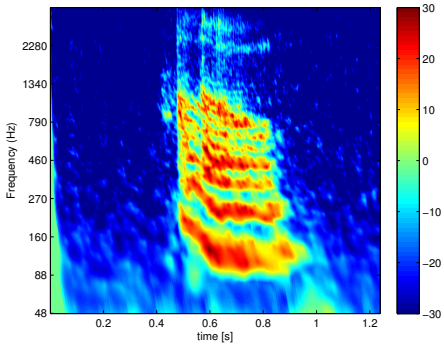
(c) Signal from figure 2.4 with pink noise added at zero dB



(d) Local target to non target ratio of the signal in figure 2.5(c)



(e) Signal from figure 2.4 with white noise added at zero dB



(f) Local target to non target ratio of the signal in figure 2.5(e)

Figure 2.5: The left column shows the clean “hallo” from figure 2.4 with several types of noise added at 0 dB SNR and the right column the local target to non target ratio of the sounds

Select regions from a cochleogram that are highly likely to stem from a single source in realistic acoustical circumstances

Chapter 3 introduces a new method for signal processing. It exploits the properties of tones and pulses to extract segments in the cochleogram that are likely to belong to a single source.

Show a possible recognition strategy for these regions

Chapter 4 shows a recognition strategy based on forming groups of regions found with the methods in chapter 3. Nearest-neighbor algorithms are used to classify these groups. Initially it is not required to be perfectly correct. However, the general pattern of sources must provide a good indication of the actual auditory scene, wherever the system is deployed.

Show integration with knowledge driven context information

As the contents of the datasets is highly ambiguous, the performance of signal-driven techniques will not be maximally high. To disambiguate some detected sound event it is coupled with a knowledge-driven network. This combination is described in chapter 4

Show performance on several different datasets

Before performance on any dataset can be shown, it is necessary to create a ground truth for the recordings being analyzed. Chapter 5 discusses this process and its problems for environmental sounds.

Part III of this thesis shows the performance of the recognition system on diverse datasets and tasks. Chapter 8 extends the grouping with a formant detection algorithm and results of vowel identification are shown. Chapter 6 shows the application to recordings from a train station with a focus on the detection of verbal aggression and related classes. Also the integration with results from video processing is discussed. Finally in chapter 7 the methods are applied to two datasets and combined with a dynamic network to supply context information. This addition improves performance. The first dataset used here is a large dataset of recording from the town of Assen. The number of classes in this dataset is high (N=54). The second is the dataset from the Amsterdam Amstel station, also used in chapter 6.

Part II

Theory of sound source recognition in uncontrolled environments

3

Tone, pulse, and chirp decomposition

This chapter is based on: Johannes D. Krijnders and Tjeerd C. Andringa. Tone, pulse, and chirp decomposition for environmental sound analysis. *Acta Acoustica*. in preparation.

This chapter introduces two efficient, robust, and informative low-level signal representations, the tone- and the pulse-fit (TF resp. PF), that are suitable for environmental sound recognition. These representations are based on a time-frequency distribution which allows a decomposition of the signal into tonal and pulse-like subsets. These subsets allow the estimation of source properties. This decomposition is part of an effort to design an environmental sound recognition (ESR, see chapter 4) system that can recognize sound events as diverse as passing cars, singing birds, opening doors, playing children, the presence of music, the sound of wind and rain, and passing aircraft. Initially the system is not required to be perfectly correct on all sound classes. However, the general pattern of recognized sources must provide a reliable indication of the actual auditory scene, wherever the system is deployed. This combination of demands poses constraints on the low level representation, which leads to the representations proposed in this chapter.

The proposed representations are part of a novel approach to environmental, every-day, or real-world sound recognition that is designed to work

reliably in the same width of acoustic environments as the mammalian auditory system. This approach matches signal-driven evidence with knowledge-driven expectation (knowledge, (Niessen et al., 2009b), chapter 4). This chapter focuses on estimating evidence about the presence of tones, pulses, and chirps. This evidence can form the basis of signal representations that are independent of the acoustical environment. The representations can be used as constituents of "auditory objects" that represent information about a single source or process. A central idea of the representations proposed in this chapter is to exploit the continuity of the development of patterns in the signal, that are imposed by the physical source that provides the signal energy (Andringa, 2002).

Because the final application is in environmental sound recognition, it is difficult to prove the validity of a low level measure since no ground truth can be determined on this level from an uncontrolled environmental sound. However, slightly broadened variants of TF and PF have been shown to form a suitable bases to define auditory textures of real-world sounds (Andringa, 2008) where TF and TP were referred to as tonality and pulsality). These textures were applied to a database compiled by Gygi (Gygi et al., 2007), which contained 100 different non-speech sounds recorded in many different (and unknown) situations. The study showed that a TF- and TP-based method was able to explain much of the perceptual difference between very different environmental sounds. Therefore, we focus on the technical validation of the properties of the tone- and the pulse-fit.

Compared to other sound recognition tasks, ESR poses a number of different demands. In many pattern recognition tasks it is not unreasonable to demand that the input consists of a single source class. For example, in speech recognition the input can be assumed to be speech in a known language that is produced by a cooperative speaker and has minimal, or known and stable, transmission effects. Therefore it is effective to use features, like Mel-frequency cepstral coefficients (MFCC's, which were developed for speech recognition with Hidden Markov Models(HMM)). However, in the case of environmental sound recognition there is no such thing as a "cooperative source". In fact, the pattern of sources has to be estimated from the signal, since no restrictions on the signals content, other than that it stems from a superposition of physical sources, can be assumed safely. Additionally, all sound sources are influenced by varying, and typically unknown degrees of transmission effects; especially in the form of reflections that add delayed copies of the source signal to the input.

Another important difference is that both the input and the output of a speech recognition system are ordered developments (albeit in very different domains). This constrains and therefore facilitates decoding and is essential for the ASR design. In contrast, in ESR the individual sound sources may be either uncorrelated or subject to complex within-class and between-class correlations. Moreover, different sound sources can be defined on quite different

temporal scales. The result is a varying superposition of sources instead of a sequence of events with a temporal ordering that is reliable enough to guide decoding. Finally, sounds from more distant or more diffuse sources tend to merge and form a changing diffuse background that might be quite different from the sources that constitute it. Unlike in automatic speech recognition (ASR) this background is informative and needs to be described as well.

Humans have little trouble to solve this targets-in-noise problem (for the specific case of speech generally known as the cocktail-party effect (Cherry, 1953; Cherry and Taylor, 1954; Bronkhorst, 2000)). Computer implementations on the other hand have problems when speech is mixed with low levels non-stationary background sounds (Gong, 1995; Lippmann, 1997; O'Shaughnessy, 2008). One of the problems of current sound recognition systems is that the MFCC or similar features work best in clean conditions. Concurrent sources influence all coefficients in unpredictable ways. As long as the other sources can be treated as a small perturbation of the target signal this may work, but in many cases this cannot be guaranteed. The unpredictability of the perturbations makes it hard to separate concurrent sources from the target sound and to properly recognize the target (O'Shaughnessy, 2008).

Human performance seems to benefit from the combination of signal-driven processing and top-down knowledge (Shinn-Cunningham, 2008). To mimic such an approach both processes should share common representations and have the possibility to reason about multiple interpretation hypotheses. These hypotheses could relate to the auditory objects that appear in modern cognitive research (Shinn-Cunningham, 2008) . What an auditory object is is still unclear (Carlyon et al., 2001; Griffiths and Warren, 2004), but it should represent information from a single source. The tone-fit and pulse-fit measures introduced in this chapter help to select time-frequency regions that are likely to stem from a single source, which is to be contrasted to approaches were the complete scene is identified as a whole (Aucouturier et al., 2007; Chu et al., 2009; Eronen et al., 2006). However, these methods can activate top-down knowledge to disambiguate the sound sources found with the proposed methods (see also chapter 4).

The chapter continues with a background of computational auditory scene analysis (CASA) methods. It uses a number of results, more than fifty year old (Allen, 1994; Cherry, 1953), as basis for a set of demands on the tone- and the pulse-fit. The methods section describes the auditory model and the TF and PF algorithms. The experiment section describes the experiments that demonstrate that the tone- and pulse-fit algorithms satisfy the demands and show their application to real-world sounds. Finally the chapter discusses the results, and it touches on a possible application to speed up sparse coding.

3.1 Background

3.1.1 ASA and Auditory Objects

The main difference between human perception and modern ASR performance might be related to a difference in signal representation. As a number of decades of research on Auditory Scene Analysis (ASA) have shown (Bregman, 1990), humans are able to track the development of (patterns of) signal components such as tones and pulses. The important characteristic of these patterns is that it allows the auditory system to form interpretation hypotheses that are very likely to stem from a single source (Bregman, 1990). Unlike the name of the research domain suggest, ASA has not often been aimed directly at real auditory scenes. Instead ASA has shown which rules govern the grouping of evidence into perceptual streams by focussing on basic patterns of tones, noises, and clicks. This streaming behavior has been modeled (Wang and Brown, 2006; Ellis, 1996), and these implementations have solved some problems in sound source separation and robustness to noise. However, these implementations typically involve re-synthesizing audio from a selection, called a mask, and using this as input for a standard ASR system. This extra step requires a hard (yes/no) decision on what to include in the re-synthesized sound before the signal is actually recognized and positively identified as target. This strict early-state selection may be suboptimal because it is error-prone. A related approach, called missing data theory (Cooke et al., 2001; Cooke, 2006), accounts for the fact that some regions of the time-frequency plane might be more important than other regions and ought to be weighted differently in the decoding process. Both approaches involve the estimation of a mask of which low-level signal properties indicate that it is more likely to represent target than non-target.

3.1.2 The local signal-to-noise ratio

As noted by Haykin (Haykin and Chen, 2005), the cocktail-party phenomenon is, fifty years after Cherry's seminal work, still an enigma and the "answer to the cocktail-party phenomenon requires deep understanding of many fundamental issues that are deemed to be of theoretical and technical importance." The cocktail-party phenomenon can be described as the ability to detect and recognize target sounds that are mixed with and partially masked by similar sounds. So even when the target does not stand out in terms of energy or spectral content, i.e. it is non-salient, it can be detected and recognized. Natural pattern detection and recognition can also rely on more subtle cues than saliency. As summarized in (Haykin and Chen, 2005, based on Bregman (1990)) human auditory scene analysis relies the estimation of coherent units of single-source evidence, time-frequency elements or signal components, in combination with a number of principles to group these signal

components. This chapter focusses on the estimation of signal components that are narrow in the time-frequency plane: tones, pulses, and chirps. We will term these "narrow signal components".

Signal component estimation can be coupled to a sixty year old result by Fletcher (Fletcher (1950), reviewed in (Allen, 1994)) who has determined the local signal-to-noise ratio (SNR) as a necessary and sufficient indicator of the reliability of acoustic evidence: time-frequency regions with a negative local SNR (in dB) did not contribute to recognition performance, but a positive local SNR improved phoneme recognition. A local SNR exceeding 30 dB did not lead to further improvement. The combination of signal component estimation and the notion that all regions with a positive local SNR should be able to contribute to the probability of a correct recognition result, forms the basis of this work. In situations where the signal-driven evidence is reliable it will lead to the activation of interpretation hypotheses for possible groupings. When the signal-driven evidence is less reliable, grouping hypotheses based on context, task demands, or prior knowledge can use the less reliable evidence to decide on the best grouping of signal components. The resulting pattern of grouped signal components provides information about the development of the source that produced it.

In this chapter we focus on the reliability of signal component estimation in terms of the local signal-to-noise-ratio (SNR). We will relate this to Fletcher's early findings in phoneme recognition, which is a related but qualitatively different task. For ASA the SNR of tones and pulses is not important as signal property. What is important are the properties of the individual signal components and especially which fraction of it can be estimated. This deteriorates as function of decreasing SNR. In favorable situations, narrow signal components can be found and tracked by stringing spectro-temporal peaks together. When broadband background sounds become progressively stronger and start to approach the energy of the narrow signal components, the influence of the "noise" in the estimated signal components becomes more and more pronounced. Initially this leads to a slight modulation of the energy of the narrow signal components. The modulations in energy and peak position increase further until the modulations become so pronounced that peak-tracks break up. Even stronger background sounds masks the target more and more until the target is completely masked and the background is locally dominant (although it might be possible to find it with more global measures such as an autocorrelation (Krijnders et al., 2007)). The left panels of figure 3.2 show the deterioration due to local SNR decrease. The upper left panel shows a cochleogram of a sinusoid. The lower left panel shows a sinusoid with a decreasing SNR and a number of signal components as strings of neighboring peaks. With decreasing SNR ratio the string breaks up into shorter parts. However, the reliable parts can be used to generate expectations that justify a reconnection or less reliable parts well beyond the first break-up. Furthermore, when the context provides a matching pitch

contour, the system can, ideally, switch from an orienting mode, in which the signal drives processing, to a checking-mode in which knowledge-driven hypotheses search for matching evidence. In the orienting mode the signal must be sufficiently unambiguous to drive processing. In the checking mode, knowledge and expectations counteract signal ambiguity, which allows the system to function in more (adverse) situations. We suggest that the interplay between orienting and checking modes of pattern recognition are an important, and hitherto, neglected ingredient of the cocktail party phenomenon.

3.1.3 Required properties

This chapter aims to provide a low-level signal representation that can be combined with knowledge-driven expectations through a well defined relation between local spectro-temporal patterns and the local SNR as a measure of (statistical) reliability. In benign conditions the tone-fit and pulse-fit should facilitate signal-driven interpretation hypothesis activation, while in more challenging conditions, or when a suitable interpretation hypothesis is available, it should provide information in a form that can easily be checked for consistency with the expectation. Therefore the resulting representations should have the following properties:

1. ability to detect all narrow signal components, tones/pulses/chirps, with a high local SNR,
2. sensitivity to noise in a predictable and local SNR dependent manner,
3. level independency (i.e. only local SNR dependent),
4. continuity preservation: a smooth (continuous and continuous first derivative) development through time and frequency should lead to a single signal component with a similar development,
5. frequency independency,
6. correspondency to a measure of local SNR, and
7. ability to work not only on laboratory grade pulses and tones, but also on a wide range of minimally pre-specified environmental sounds.
8. 100% detection score on true tones, pulses and chirps.

3.2 Methods

Our methods share a common ground with many existing CASA systems in that the TF processing is performed with a cochlea model. Matched filters compute the tone- and pulse-fit.

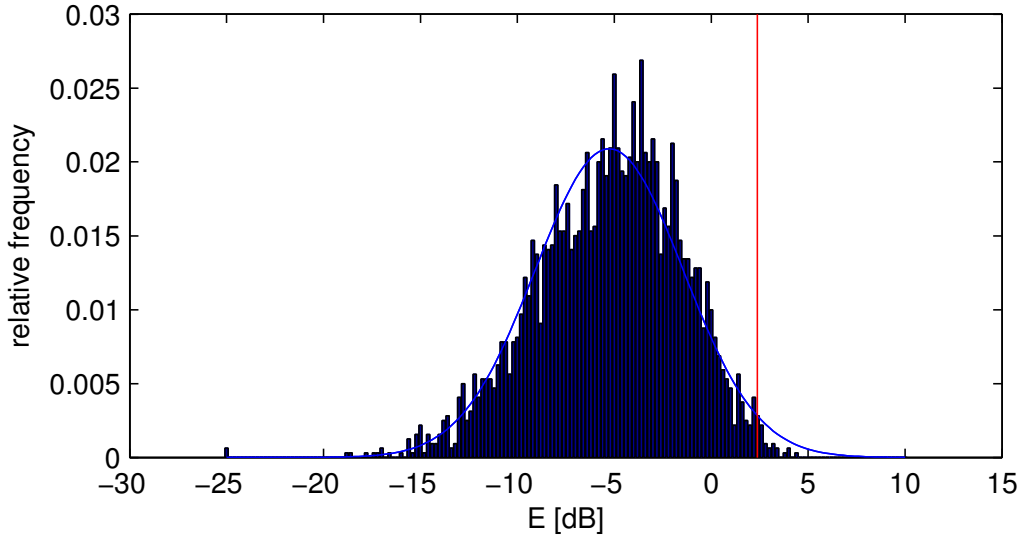


Figure 3.1: Distribution of the energy in channel 50 (histogram) and the best fitting gaussian distribution (line). The red line indicates the $2\sigma_n$ point, only 2.5% of the energy exceed this value. The absolute value of E is arbitrary.

3.2.1 Time-frequency processing

The TF processing is performed by the gamma-chirp filterbank as described in 2.3. The filterbank has 100 channels and a maximum frequency of 4000 Hz.

3.2.2 Standard deviation of noise

While tones and pulses lead to an excitation with predictable shape, broadband noise gives a cochleogram excitation pattern that does not repeat itself and that can only be represented with a probability density (pdf) function. This pdf approximates a Gaussian (Figure 3.1) that can be described well with a segment dependent mean, representing the average energy, and a standard deviation σ_n, where n denotes the channel or segment number. The standard deviation represents the spread of the noisy fluctuations that constitute the inevitable fine-structure of aperiodic contributions. We will use σ_n to normalize the local signal-to-noise ratio with the local standard deviation. The local standard deviation is estimated by exciting all segments with white noise. Given the near Gaussian energy distribution of the energy fluctuations this entails that about 2.5% of the fluctuations exceed 2 standard deviations above the mean. The scale-invariances of the system ensure that both the level and the overall shape of the reference noise are unimportant as long as its average is locally flat.

3.2.3 Tone and pulse fit

To detect tones and pulses in the cochleogram, we use segment dependent matched filters that provide a measure of fit with an ideal pulse- or tone-shape for all cochleogram channels n. These matched filters are derived from the cochleogram response to ideal tones and pulses. The algorithm for creating and applying these filters is illustrated in figure 3.2. The filter is channel dependent, so the filter definition process is repeated with tones selected to peak at each segment according to the place-frequency relation. For each segment the energy at an adjustable threshold value $th_n\sigma_n$ under the peak is determined, typically $th_n = 2$ for all segments. The sine-broadness consists of two numbers, since the response of the cochlea to a tone is asymmetric. The differences between the position of the top and the upward and the downward flank at the threshold value constitute the sine-broadness $sb = sb_1 + sb_2$.

The tone-fit of the segment is the difference between the energy of the center minus the mean of the energy of one sine-broadness before and after the center. This difference is normalized by the local noise standard deviation σ_n:

$$TF = \frac{E(n) - \frac{1}{2}(E(n - sb_1) + E(n + sb_2))}{th_n\sigma_n} \tag{3.1}$$

The TF can only be computed for channels around which a sufficient scope is available. This entails that the TF has less desirable properties for high and low frequencies (edge effects). This can be reduced by calculating more channels that strictly necessary.

The pulse-fit uses the same algorithm, but in the temporal direction instead of the frequency direction. The filters lead to the highest values for ideal excitations (which show the sharpest possible local response), but react to non-ideal tones and pulses with lower values as well. In the case of chirps this entails that both the tone fit and the pulse fit contribute. The steepness of the chirp determines the relative contribution (see figure 3.4).

3.2.4 Separating noisy fluctuations from “real” signal contributions

The resulting measures corresponds to a segment independent measure for the local SNR in terms of the local standard deviation of noise. It can be interpreted directly as a reliability of tonal and pulsal evidence. However, the fluctuations of noise can be locally similar to tones, pulses, and especially (down-) chirps. These random look-a-like contributions are unlikely to last very long and have rarely considerably more energy (although this is statistically possible) than their immediate environment. Therefore, these noisy fluctuations can be separated from “real” signal contributions by assuming

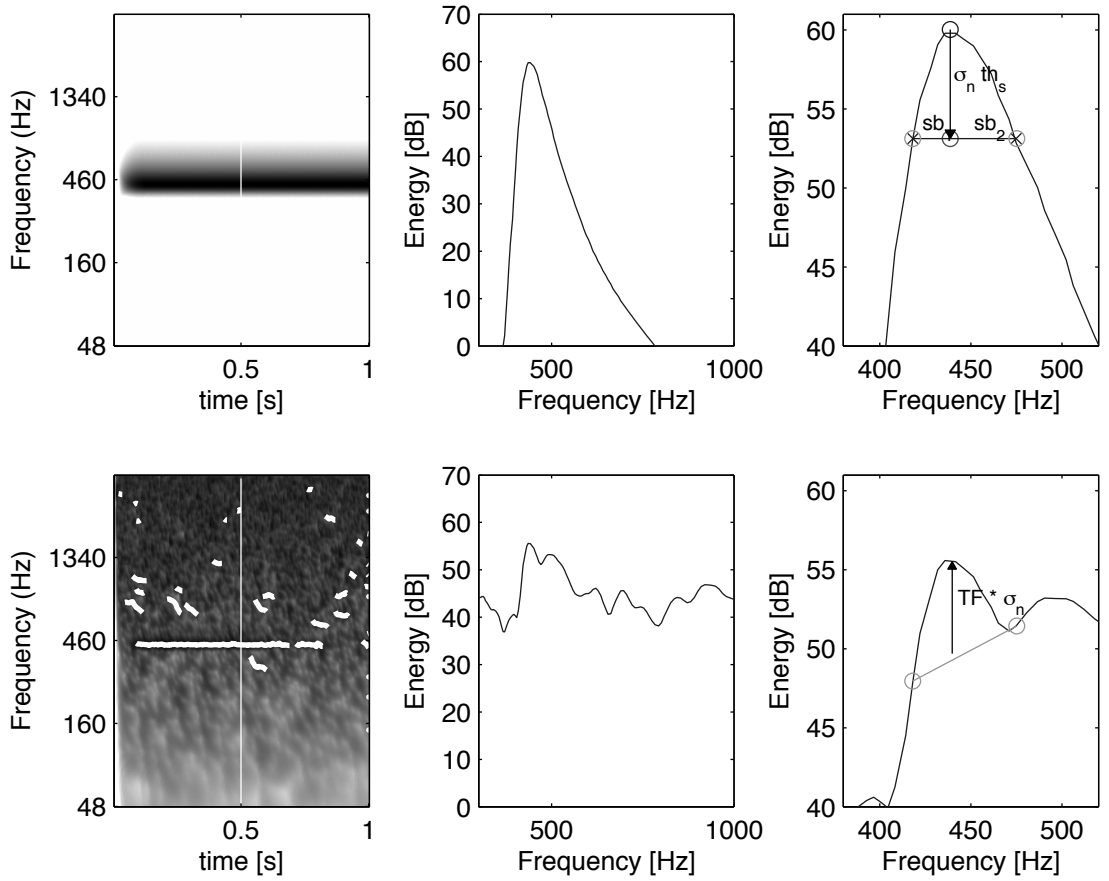


Figure 3.2: Computation of the tone-fit (TF). The upper left panel shows an ideal sinusoid, the lower a noisy sinusoid with a decreasing SNR and the extracted signal components, including some spurious ones. The upper middle panel shows an ideal sinusoid response around the peak (in the segment direction). The lower middle panel shows the cross section around the noisy pulse. The upper right panel shows the computation of the TF at the peak position. The TF is the energy difference denoted by the vertical line. The lower right panels shows the TF computation for the noisy tone. The computation of the pulse-fit is similar, but with frequency replaced by time.

that their duration is too short (for tones) or that they do not span enough segments. To compute this measure a connected components (van der Heijden, 1994) algorithm was applied to all regions with TF or TP values exceeding the threshold of 2σ. This size threshold was chosen so that 95% of the connected components in white noise are classified as noise and only 5% of the regions is classified as non-noise. Because the threshold of 2σ already discards 95%, the resulting area incorrectly assigned to non-noise is considerably less than 5% of the total area.

Additional higher levels of processing should deal with the remaining spurious contributions. Conversely, some of the target information near the local SNR threshold is incorrectly discarded. However, this part of the target was not very reliable in the first place. Furthermore, these regions can also be re-evaluated by higher levels of processing. Since the reliable evidence is never influenced by this procedure, the net effect of this procedure is a strong bias towards more reliable signal-driven evidence.

3.3 Experiments

To demonstrate some of properties of the tone- and pulse-fit we provide the results of a number of experiments. All experiments where performed with the cochleogram calculation described above. The cochleogram has a maximum frequency of 4000 Hz. The first section compares the use of thresholded TF values to optimally thresholded energy levels. The TF values and the optimal threshold are shown to be completely energy level independent, while the energy threshold is strongly dependent on energy. The section 3.3.2 shows how the TF and PF values change for chirps with different chirp speeds. As the chirp speed increases the TF value reduces and the PF value increases, thus making it possible the estimate the chirp speed based on the combination of both measures. The third section shows that the TF values have a well defined correlation with the local target-to-noise ratio. The fourth section shows how the fraction of a pure tone that can be detected changes with increasing noise levels. The fifth section show the accuracy of the tone-fit and pulse-fit. The accuracy for the tone-fit is better than than 0.8% for most frequencies, while the pulse-fit is accurate with 2 ms, with a frame-size of 5 ms. The sixth section shows that two tones can be separated based on tone-fit if they differ at least 3% in frequency. The seventh and final section shows the behavior of the tone-fit and pulse-fit on isolated recordings of everyday sounds from the Gygi database Gygi et al. (2007). The recordings in the dataset are labeled as being harmonic, pulse-like or continuous sounds and the tone-fit is shown to explain most of the energy in the harmonic class and little in the two other, whereas the pulse-fit explains most of the energy in the pulse-like class and little in the other two.

3.3.1 A comparison with energy threshold

Standard methods to estimate a target in noise rely heavily of the estimation of a noise level. This noise level has to be estimated from low-level signal properties (typically energy or harmonicity) before the signal is recognized and can be known what is target and and what "noise". The noise level is typically adapted dynamically with a time constant and some form of temporal averaging (Martin, 2001; Rangachari and Loizou, 2006). While this works for many cases, the noise level estimation may not be optimal or even quite wrong because it is based on an incorrect decision on what is target and what is "noise". We show that the tone-fit measure is able to perform the target/noise estimation task without the need to estimate a noise-level. We only assume a narrow signal component as target. In the upper panel of figure 3.3 we show that the optimal energy threshold is a function of the signal level. To calculate the optimal threshold we have defined three masks, a ground truth (M_{GT}), a energy threshold mask (M_{EdB}), and a tone-fit threshold mask (M_{TF}):

$$M_{GT} = P_n < P_s \tag{3.2}$$
$$M_{EdB} = P_{EdB} < th_{EdB} \tag{3.3}$$
$$M_{TF} = TF < th_{TF} \tag{3.4}$$

We optimize on f-measure, the harmonic mean of precision and recall (see section 4.3.1, equations 4.5, 4.6, Baeza-Yates and Ribeiro-Neto, 1999). This entails that we punish both the exclusion of target as well as the inclusion of noise in M_{EdB} and M_{TF} , M_{GT} is used as ground truth. The target is a tone in noise with a local SNR level of 15 dB. The overall signal level is gradually increased 20 dB which entails that the energy threshold th_{EdB} has to increase as well. For 10 points in time we calculate the optimal thresholds. The middle panel of figure 3.3 shows that the optimal energy threshold scales linearly with the signal level, while the threshold for the tone-fit is constant at about 1.5 standard deviations. The shape of the f-measure (4.7) in the upper panel of figure 3.3 entails that even errors as small as a few dB in the energy-based noise model lead to prominent differences in the regions to be included in the mask. The maximum value of the f-measure is smaller for the tone-fit than for the energy because the tone-fit only selects the most energetic part around the center of the tone, while ignoring the rest of the shape, which is redundant.

3.3.2 Tone-fit and pulse-fit for chirps

Tones, defined by a single frequency at all points in time, and pulses, where all frequencies contribute at a single point in time, are extremes. Impacts, chirps and frequency modulated signals are very common in environmental

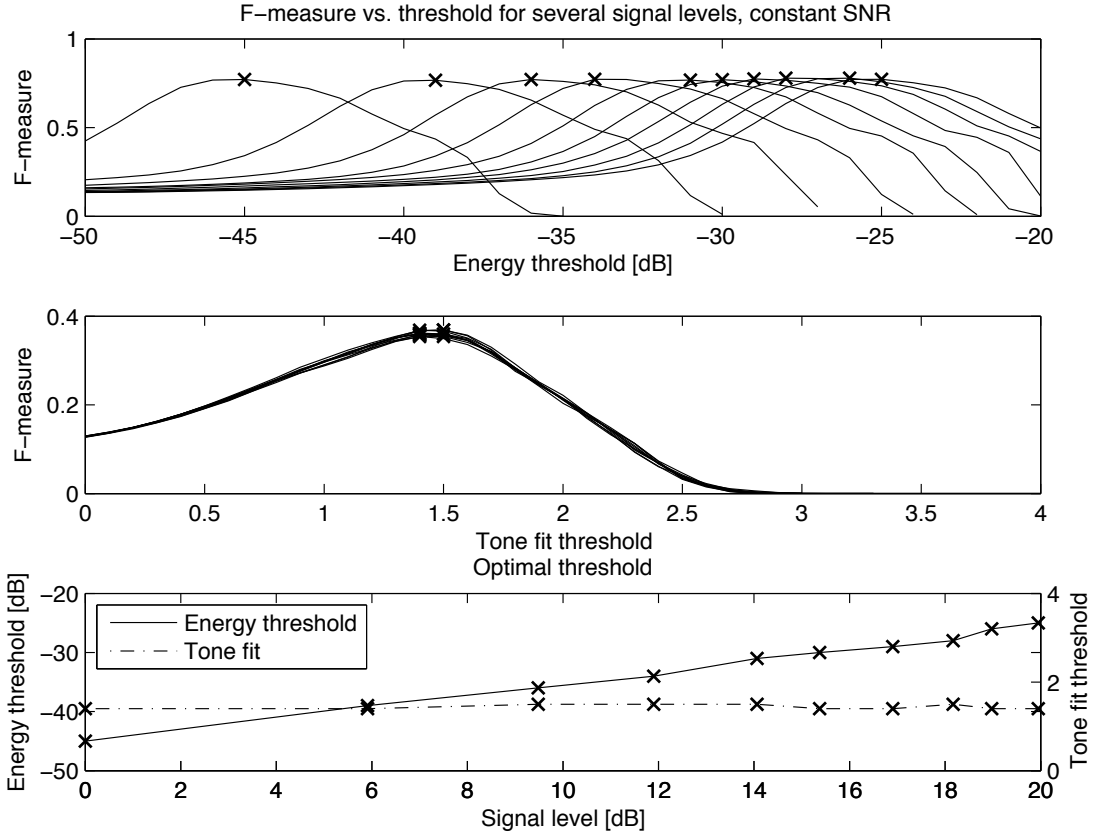


Figure 3.3: The estimation of optimal thresholds for target mask forming using a noise level estimate (upper panel) and the tone-fit (middle panel). When a signal consisting of a tone in noise is amplified gradually, the average noise level increases as well and the noise threshold should be adapted to prevent the inclusion of noise in the mask. The upper panel shows the f-measure for 10 signal levels as function of the choice of the threshold. The peaks of the f-measure follow the amplification. Off-peak values may lead to a considerable lower f-measure. The middle and lower panel demonstrates that the optimal choice of the tone-fit is independent of signal level. The lower panel summarizes the results.

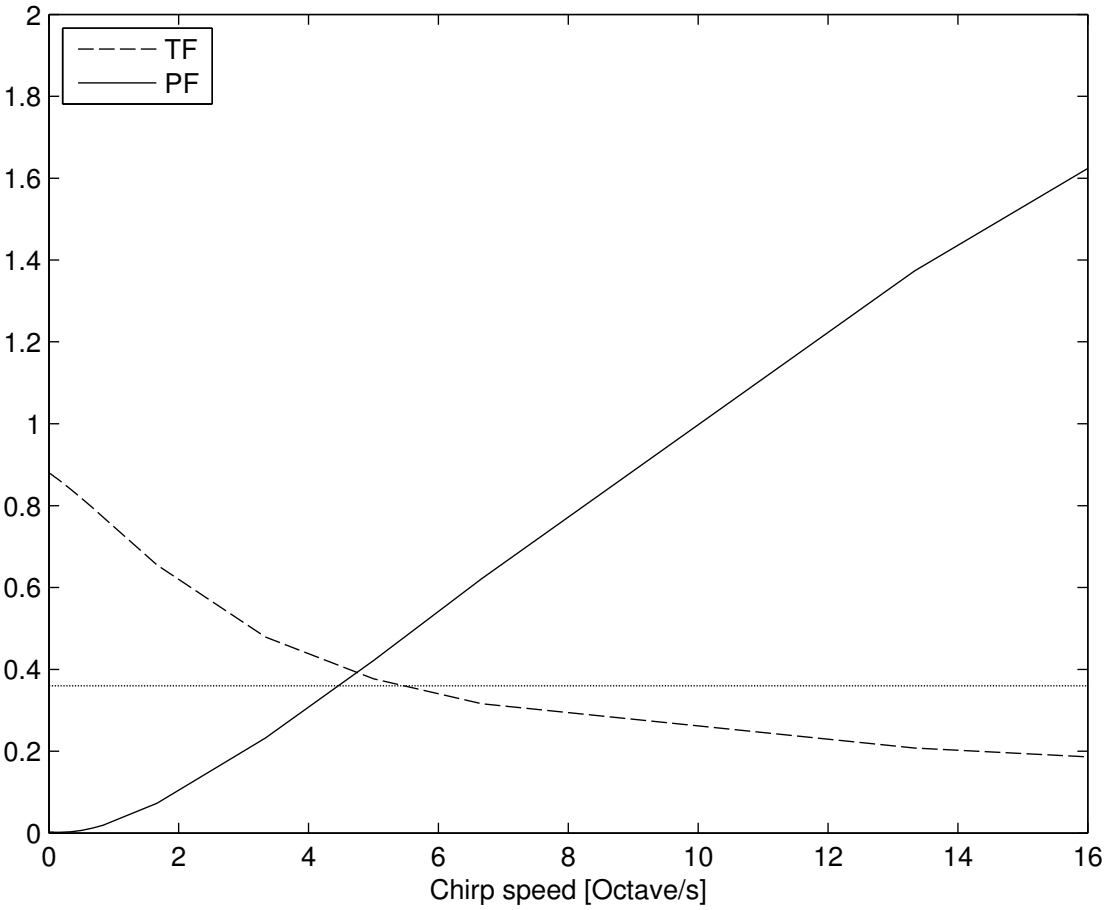


Figure 3.4: Obtained values for the tone-fit and pulse-fit measures as function of chirp steepness in octave per second. The horizontal dotted line indicates twice the standard deviation of the tone-fit for white noise. The transition from the flat tone via a chirp to a pulse has a trade-off in the values of the tone- and pulse-fit, just above twice the standard deviation of the tone-fit for white noise

sounds, such as speech, birds, cars, planes, ...) and show a non-constant instantaneous frequency (Ballas (1993), "signal", "discrete impacts", ecological frequency = 58%). Chirps can span the whole range between flat, essentially a tone, and extremely steep, essentially a pulse. It is desirable to have a well designed trade-off between the tone-fit and the pulse-fit when progressing from one extreme to the other. Figure 3.4 shows how the tone- and the pulse-fit contributions in the segment corresponding to $f_c = 600Hz$) for sweeps with increasing steepness. The chirp speed is expressed as octaves per second. Equality occurs around 5 octaves per second for the current settings of the time-frequency analysis.

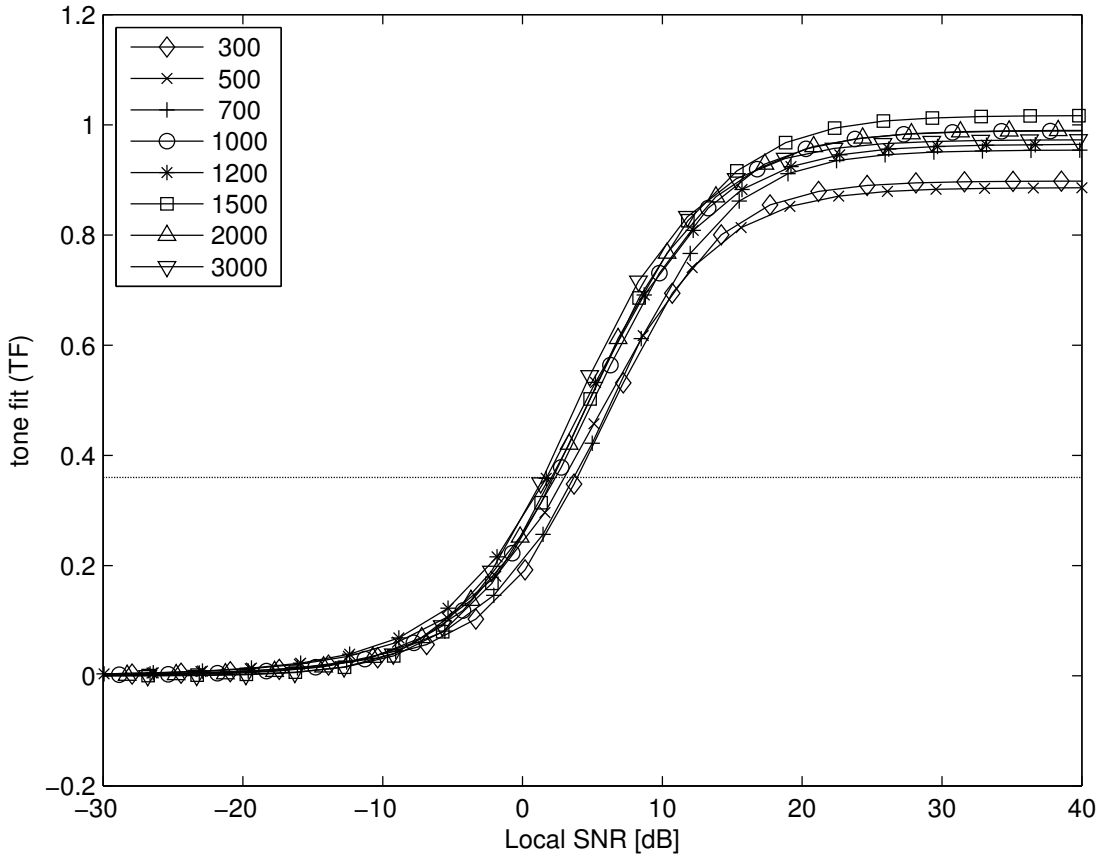


Figure 3.5: The tone-fit as function of local signal-to-noise ratio for different frequencies. The horizontal dotted line indicates twice the standard deviation of the tone-fit for white noise. This result entails that we can calculate the local signal-to-noise ratio, under the assumption that our signal is tonal, purely based on local signal properties.

3.3.3 Correlation with local signal-to-noise ratio

As stated in section 3.1.2 it is desirable to have a measure that correlates with the local SNR. In this experiment we change the noise level from -30 to 40 dB local SNR relative to a constant tone. The local SNR is calculated as the difference between the maximum energy of the tone and the mean energy of the noise in the best channel for the tone. Figure 3.5 shows the tone-fit values at the energy maximum. The deviations from one at high SNR for lower frequencies are due to numerical effects. The tone-fit correlates monotonically with the local SNR over a range of -5 dB to 25 dB which, perceptually, corresponds with range from an hardly audible tone to a dominant tone with minimal noise.

3.3.4 Tone sensitivity

As an indication of how well tones are detected and selected as function of local SNR we measured the fraction of the recovered tone duration after thresholding the tone-fit values at a range of local SNR's, expressed in σ_n. Apart from the duration-fraction we also measured the connectivity in terms of whether or not the start and end of the tone were in included in the connected component. The measures were determined for all segments between 80 and 3000 Hertz (to prevent edge effects). Figure 3.6 shows the mean and standard deviations for these two measures. The measures are by and large segment independent, with the connectivity breaking up at $4\sigma_n$ and the duration-fraction starting to decay at $2\sigma_n$. This entails that a threshold of $2\sigma_n$ is suitable for a detection strategy with knowledge-driven reconnection and a threshold of $4\sigma_n$ for a purely signal-driven analysis without knowledge-driven reconnection. This illustrates the point made in section 3.1.2 about the importance of knowledge-driven checking.

3.3.5 Spectral and temporal accuracy

This experiment shows the accuracy of frequency estimation for tones based on the TF. The frequency of a tone is changed from 80 to 4000 Hertz in steps of 20 Hertz. The estimated frequency corresponds to the point where TF = 1 on the rising slope (in frequency direction). The results (figure 3.7) show that the error is better than 0.8% for most of the frequency range. Again edge effects spoil performance for low and high frequencies. These results are sufficient for grouping of harmonics when the grouping algorithm allows for small deviations from perfect harmonic relations (Krijnders et al., 2010).

The accuracy of the PF (figure 3.8) is measured by generating pulses at random points in time and comparing the moments where the PF = 1 on the rising slope of the PF. The PF estimates the time with an accuracy of about 2 milliseconds for frequencies above 700 Hz and the maximum error is 12 milliseconds at 50 Hz. Because pulses include multiple frequencies the error in estimation of a complete pulse will be approximately 2 milliseconds. Note that these values are smaller then the frame-size of 5 ms.

3.3.6 Proximate and crossing tones

When two tones are close in frequency, two cases can be distinguished. The first occurs if the frequency difference is sufficiently large, so that the two tones result in two peaks, with an ever shallower valley in between as the frequency difference reduces. The second occurs when the tones are so close that the result is a single component with amplitude modulation due to constructive and destructive interference. The relative frequency separation

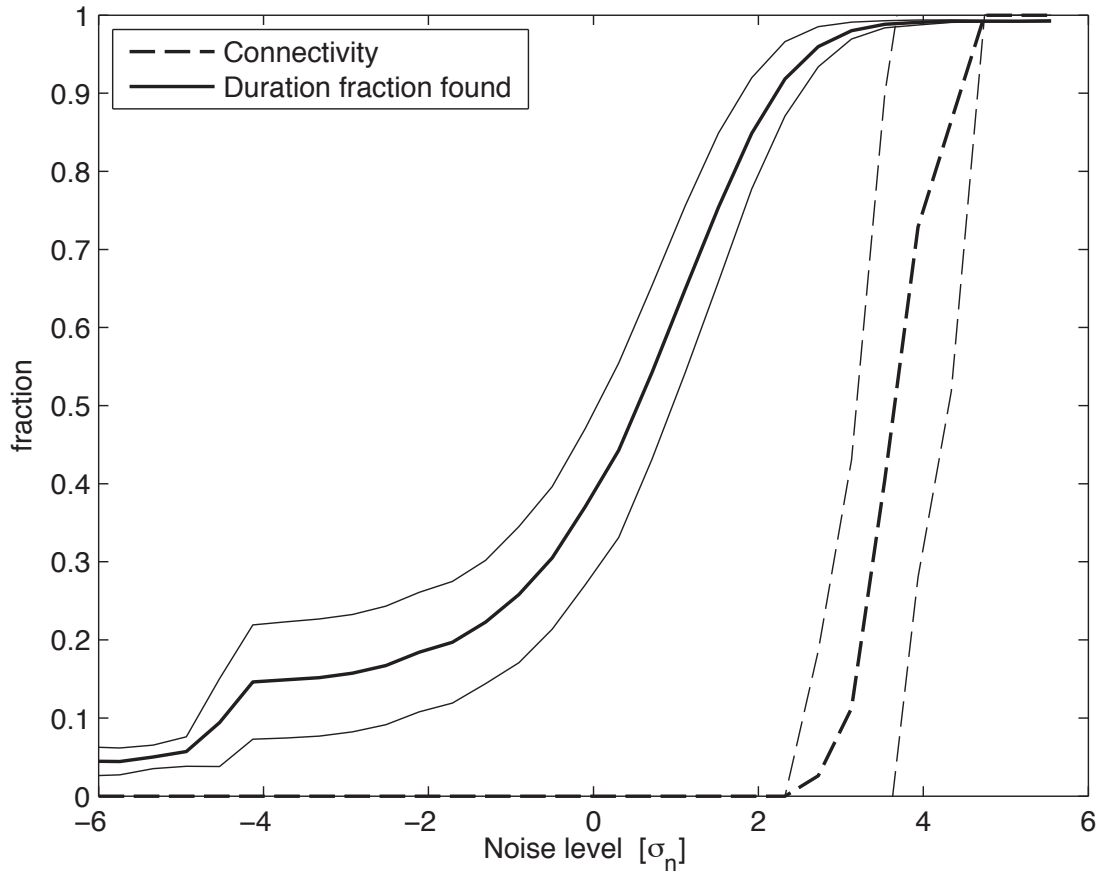


Figure 3.6: The fraction above threshold of the tone-fit of a 1 second tone as function of SNR in terms of noise standard deviations. The connectivity has the additional demand that beginning and end should belong to the same signal component and is a measure of the probability that a 1-second signal component remains a single whole in spite of the noise.

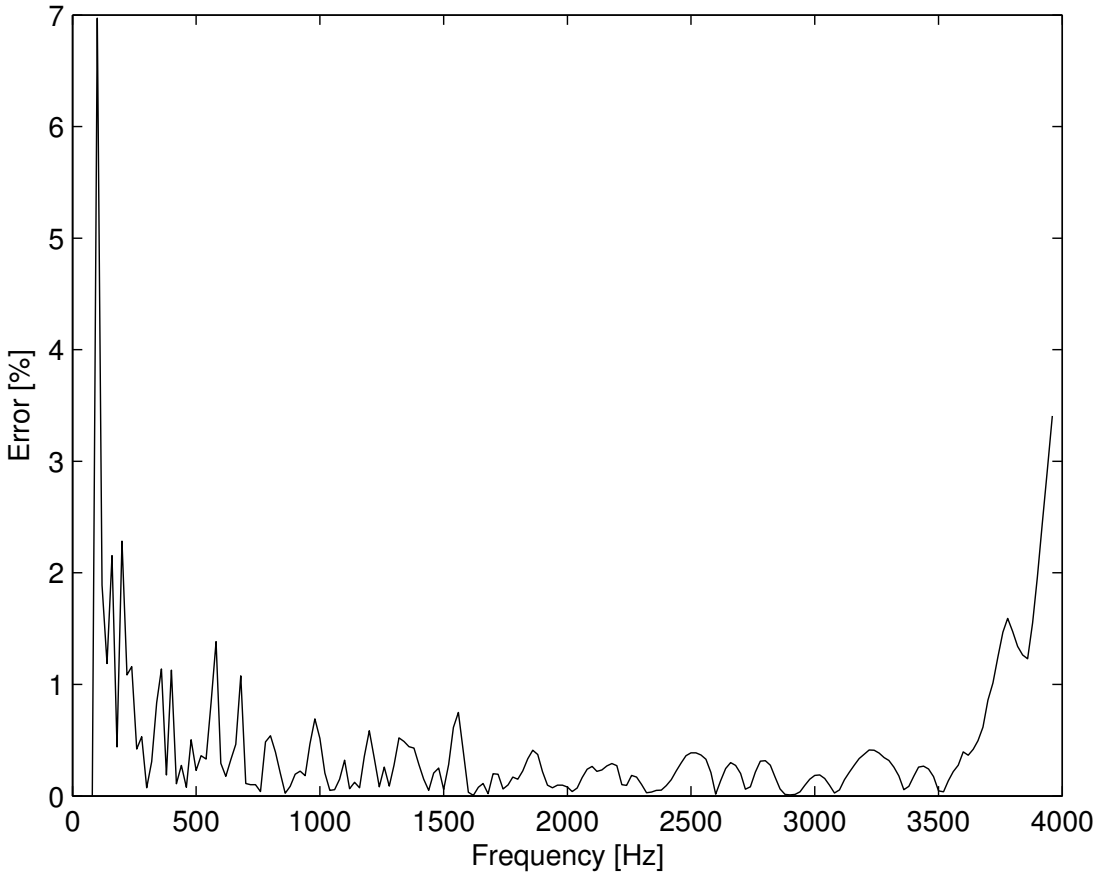


Figure 3.7: Absolute value of the relative frequency estimation error. For a large part of the frequency range the relative frequency can be estimated within 0.8%.

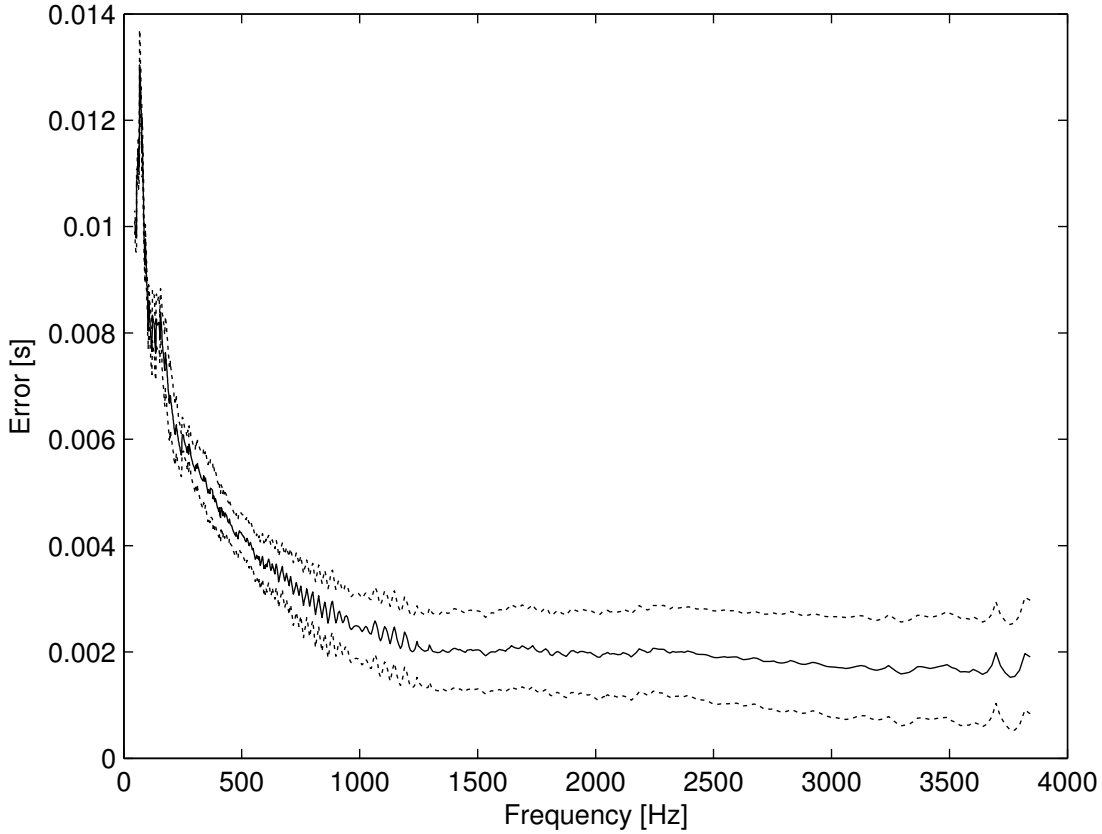


Figure 3.8: Mean and standard deviation of the estimation error in seconds. Error in low frequencies is due to a very flat slope and thus unreliable estimate. As most pulses will include higher frequencies, it is possible to estimate a pulse with an error of 2 milliseconds.

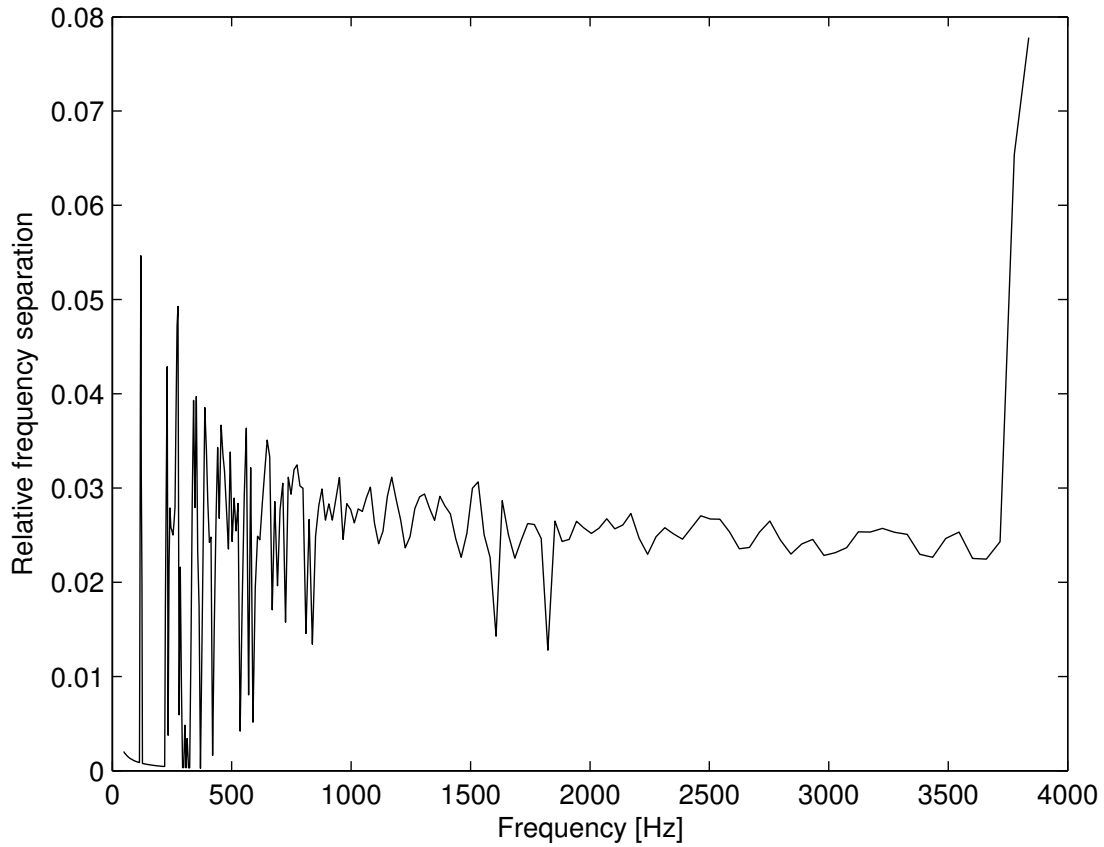


Figure 3.9: The relative frequency separation (df/f) that is just detectable with TF for proximate partials. For smaller separations the the information about the separation of the two tones resides in the energy modulation.

where the two peaks merge into a single component is plotted in figure 3.9. The TF values will follow the combined peak, because the TF values are amplitude independent. The amplitude modulation can be extracted from the energy development under the tone track. With the current settings of the TF-analysis, the minimal relative frequency separation is about 3% for most of the frequency range.

3.3.7 Proximate pulses

As with tones, when two pulses are close in time two cases can be distinguished. The first occurs when two pulses are sufficiently separated to form two peaks. The second occurs when the time difference between two pulses is so small that they merge into a single component that shows frequency modulation. The time difference for which the first case changes to the second is frequency dependent (figure 3.10). For high frequencies the resolution in the two peaks case is limited by the subsampling of the cochleogram and

$\delta t_{sep} > 1.5 dt_{fr}$. For lower frequencies the two peaks case is limited by the group delay of the filterbank.

3.3.8 Recorded sound sources

To demonstrate the effectiveness of the TF and PF measures on actual recordings of sound sources we used the database compiled by Gygi (Gygi et al., 2007). This database contains a wide range of sound sources, following Shafiro and Gygi (2004) to obtain a balanced sample of environmental sounds. The database was used to establish perceptual distances between these sounds. After multi-dimensional scaling (MDS) analysis Gygi concluded that the first two MDS-dimensions separated three classes of sounds: harmonic sounds (characterized by prominent tonal components), impact sounds (characterized by prominent pulsal contributions), and what he called "continuous sounds". The continuous sounds category contains predominantly noisy components but also some patterns of tones, pulses, and chirps. This let to the expectation that high TF-values are likely to be a good indicator of the harmonic sounds class and the high PF-values a good indicator of the pulsal sound class. For the continuous sound class the expectation was that high values of the PF and TF were not expected.

Table 3.1 shows that the fraction of the energy within the top 30dB of the cochleogram energy can be accounted for by the regions with TF or PF exceeding 1. Harmonic sounds are predominantly present in the TF, while impact sounds are predominant in the PF. Continuous sounds are not well represented in either representation. This confirms the expectation.

Figures 3.11 and 3.12 show how the mask found in clean situations degrades with a decreasing global SNR. The SNR was decreased by adding pink noise to the clean data files at different global SNRs. The rate of decay of the tone-fit based mask should be the lowest for sounds classified as harmonic, as the tonal components are likely to be the most energetic in these sounds. For impact sound the original mask should decay faster as the dominant components are not tone-like in these sounds. For continuous sounds the rate of decay should be between those of harmonic and impact sounds. For the pulse-fit measure the impact sounds should decay at the lowest rate as the SNR decreases, because in these sounds the pulse-like components are the most energetic. The rate of decay for harmonic sounds, on the other hand should be the highest. Again for the continuous sounds the rate of decay should be between the two other classes. Figures 3.11 and 3.12 do indeed show this, for the target class we find a mask with an f-measure around 0.5 at zero dB global SNR.

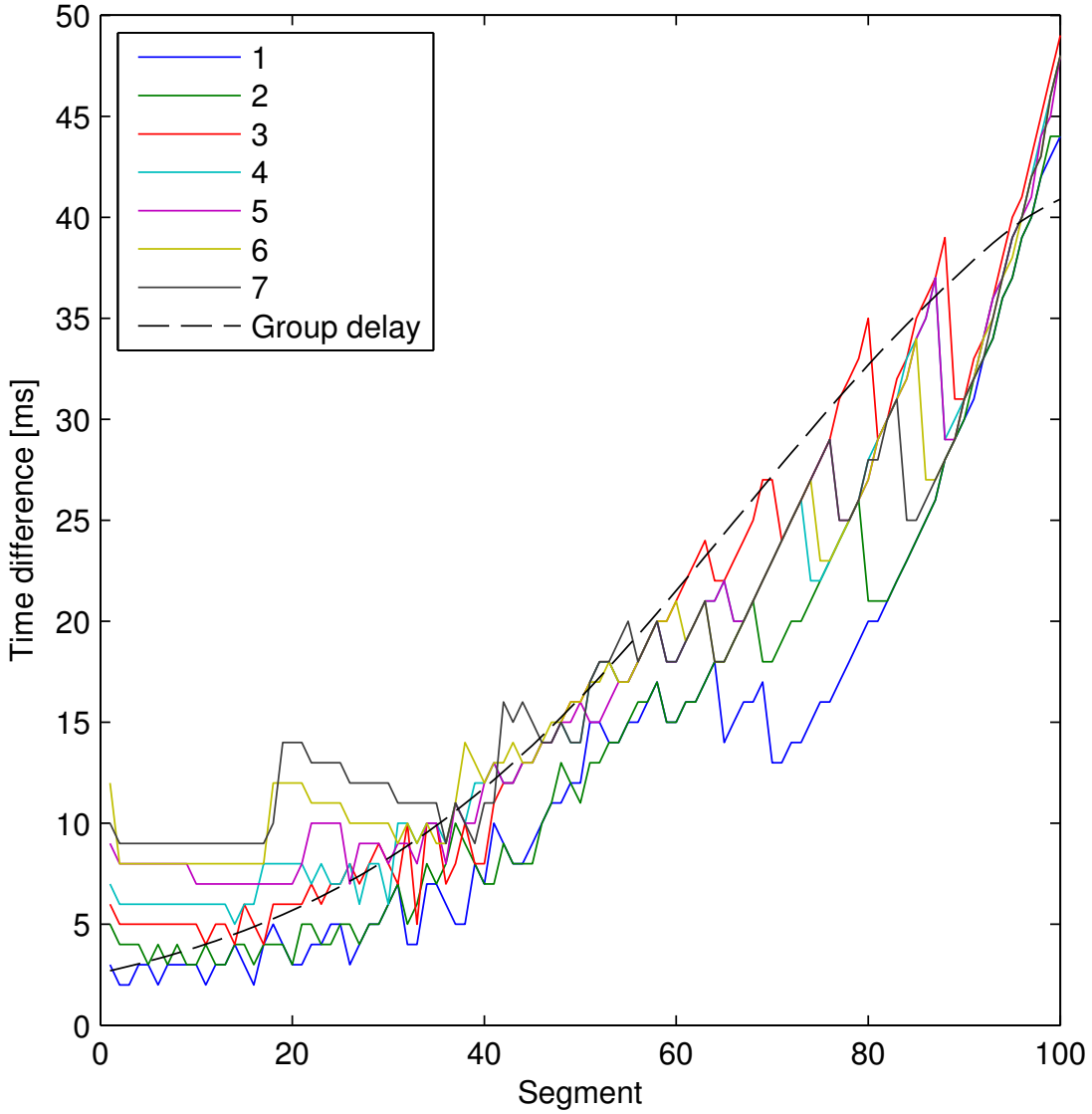


Figure 3.10: The time separation that is just detectable with PF for proximate pulses. For smaller separations the the information about the separation of the two pulses resides in the frequency modulation. The jacked lines in the low frequency (high segment numbers) are due to the frequency modulation when the peaks do not separate.

Table 3.1: The percentage of top 30 dB energy explained (mean $\pm$ standard deviation).

Measure	Harmonic	Impact	Continuous
TF	34.5 ± 12.8	8.0 ± 8.3	10.2 ± 7.5
PF	7.1 ± 8.5	29.2 ± 12.4	10.5 ± 10.3

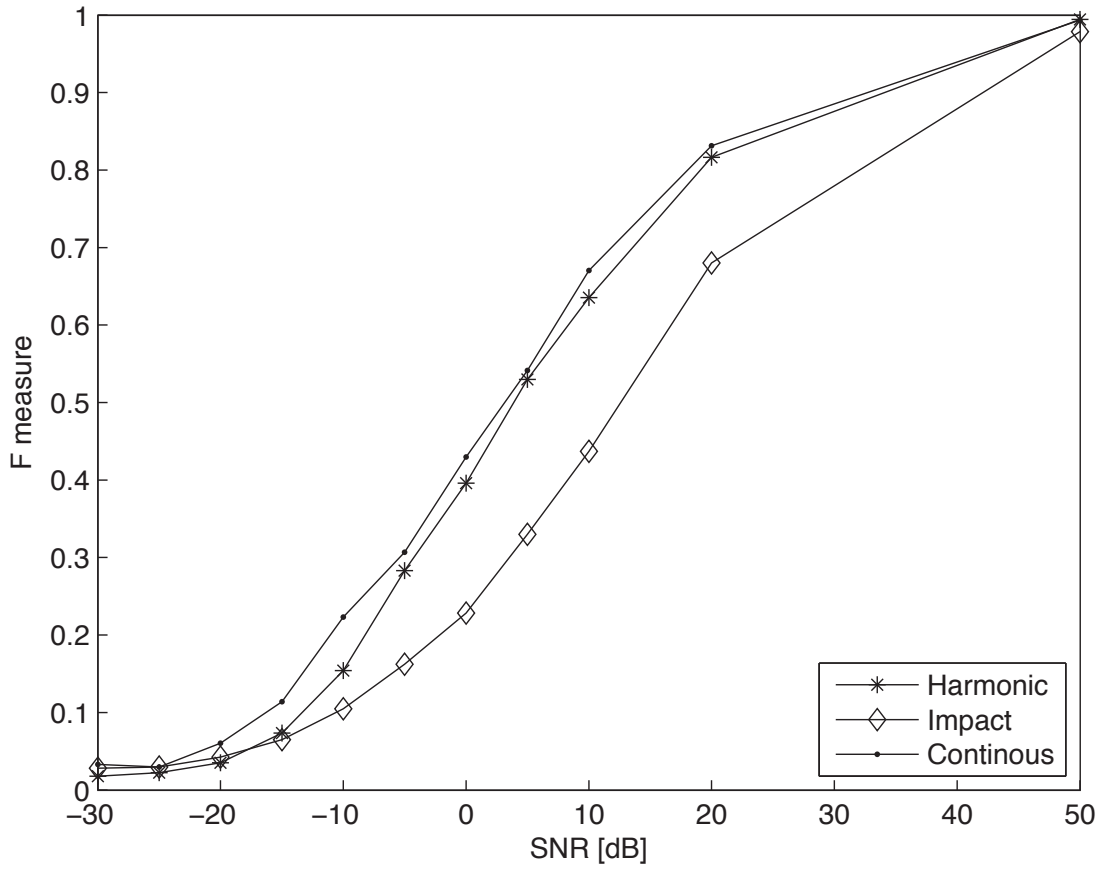


Figure 3.11: The TF decays gradually as the global SNR decreases. For impact sounds this decay is stronger than for harmonic and continuous sounds.

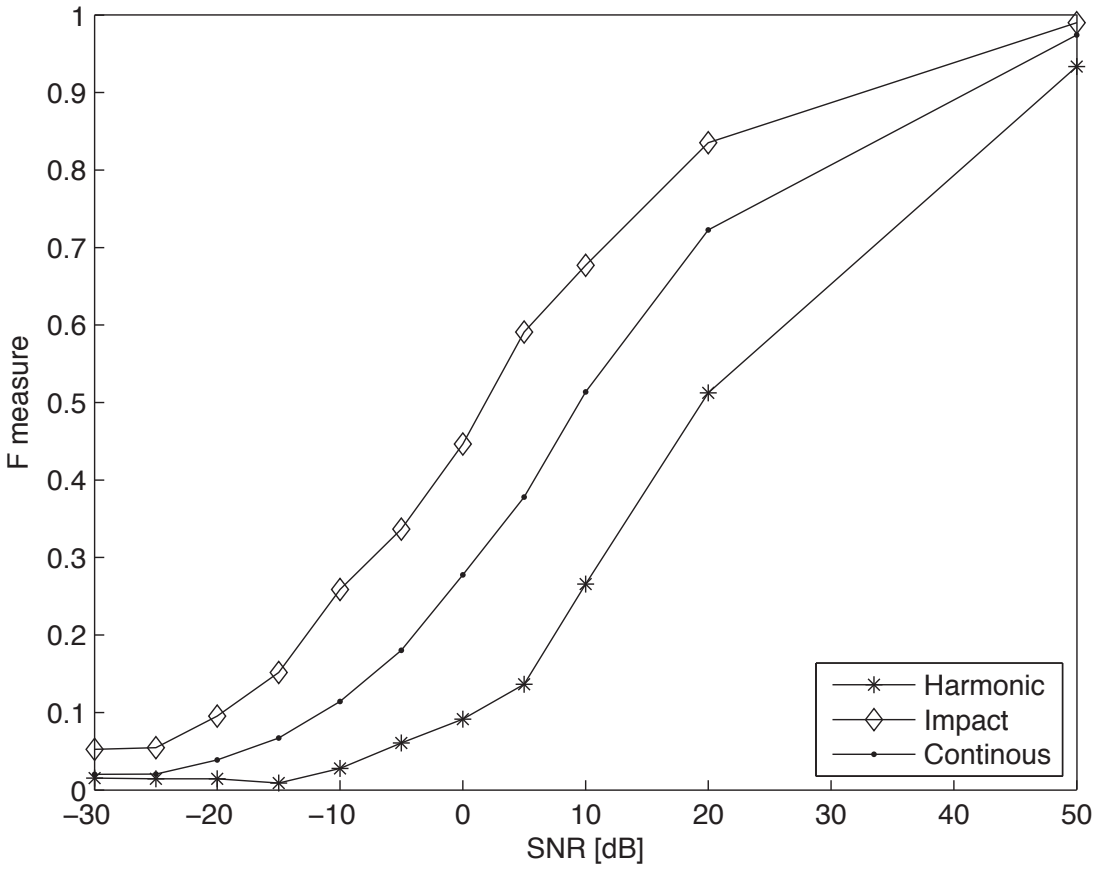


Figure 3.12: The PF decays gradually as the global SNR decreases. For continuous and harmonic sounds this decay is stronger than for impact sounds.

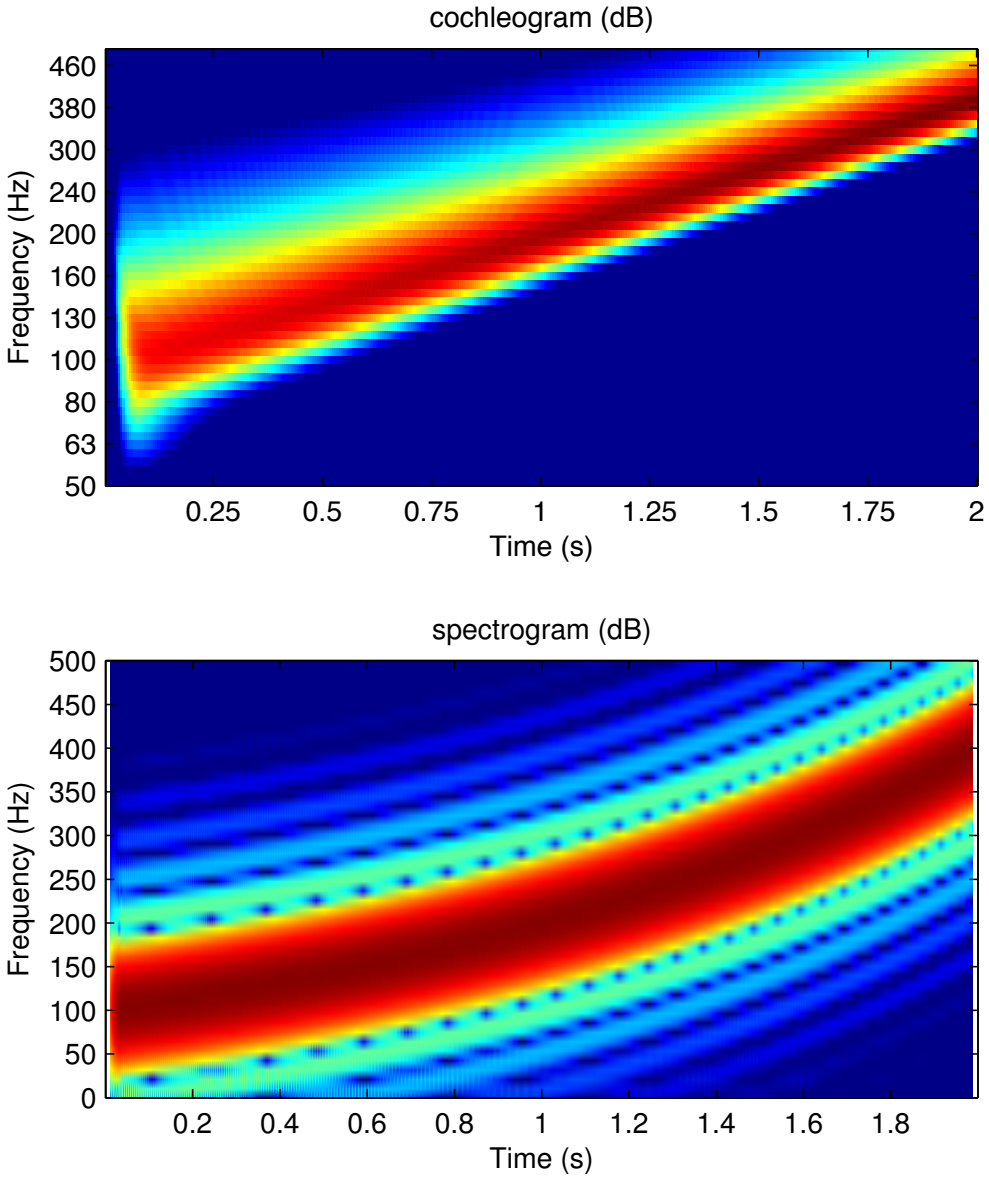


Figure 3.13: Cochleogram(top) and spectrogram(bottom) of a chirp (1 octave/s). The cochleogram is made with 100 channels and a maximum frequency of 4000 Hz. The spectrogram is made using a SFFT with a hanning window of 1024 samples (23.2 ms), an overlap of 1014 samples (99%) and an FFT length of 4096.

3.4 Discussion

The experiments have shown that it is possible to compute, detect, and describe narrow signal components conform the demands listed in 3.1.3. The algorithm relies on the predictability of the shape of pulse- and tone-responses. A highly predictable shape is not possible with regular FFT-based methods, in which signal content identical to one of the basis functions can be represented by a single basis frequency, while signal content between two basis function will be represented by a range of basis frequencies (a phenomenon called spectral leakage). Consequently, a sine sweep appears to change in spectral broadness, depending on whether or not the frequency is close to a basis frequency, see figure 3.13. Suitable windowing mitigates this effect, but it will not prevent it. We used a cochleogram derived on the basis of a gamma-chirp filterbank of overlapping filters. The gamma-chirp filterbank is not the only possible representation. Other sufficiently smooth banks of overlapping filter will work as well, such as a time domain cochlea model (Duifhuis et al., 1985). The key demand is that a smooth development at the source must be represented as a smooth spectro-temporal development in the signal representation.

The computational complexity of the tone- and pulse-fit is low and scales linearly with the number of channels. In the current real-time system, the computational bottleneck is the computation of the filterbank. In normal operational conditions the tone- and pulse-fit add only 1% - 2% percent to the processing power required to estimate the cochleogram. The parameters that define the shape and the noise statistics must be estimated offline. The offline computation of the standard deviation of white noise requires 20-60 s of noise to reach a sufficiently reliable estimate. Re-computation of TF and PF parameters is required after all changes that influence the way the time-frequency information is expressed.

3.4.1 Broadband residue

This chapter focuses on tones, pulses, and chirps that lead to a narrow expression of energy in time-frequency representations. However, broader components such as bursts, band-limited noise, and broadband noise are also important for sources like cars, whispered speech, and wind. Furthermore, the background of almost all natural sounds consists of indistinguishable contributions from many uncorrelated distant sources that lead to a broadband contribution. This broadness, in time and frequency, is characterized by random energy fluctuations.

The proposed techniques can be used to detect and characterize the centers of the narrow components. The smooth flanks of tones, chirps, and pulses will not be selected due to the focus on peaks. However, since the shape of these components is predictable, they can, in principle, be se-

lected when the development of the peaks is known. The remaining spectro-temporal area, characterized by random fluctuations with a frequency dependent standard deviation, consists of broadband or (indistinguishable) background sources. While tones, chirps, and pulses can easily be strung together to form larger units, the way to combine areas with noise-like evidence in single source representations is not directly apparent and is a topic of future research.

3.4.2 Relation with sparse modeling techniques

The purpose of the introduced measure is quite different from sparse modeling techniques (Davies and Daudet, 2006), such as (Molecular) Matching Pursuit. These methods typically optimize represented signal energy instead of reliable single source information. However, in many situations the final result, interpreted as time-frequency information about sources, may appear similar. Sparse modeling techniques approximate a signal in the time-domain by representing it as a weighted superposition of elementary waveforms from a fixed dictionary. The aim is typically to approximate the (audible) signal energy with a minimal number of contributions. As such these modeling techniques represent all sources without any physical constraints such as representing an underlying physical continuity of the source. Nevertheless the distribution of time-frequency centers of the elementary units may reflect a strong correlation with source information. The Molecular Matching Pursuit (Daudet, 2006) is a variant that constraints the selection of elementary waveforms by requiring them to form larger units according to a source model (typically certain types of music). If the source model is suitable, this provides considerable advantages and it leads to units that can be interpreted in terms of source properties.

However, all sparse modeling techniques require considerable processing. They operate in the time domain with a high sample rate and represent a computational load of the order $Nlog(N)$, where N is the dictionary size. This makes the application of these techniques less suitable to determine time-frequency regions where well-formed and reliable tones, pulses and chirps might exist. Although the tone- and the pulse-fit have no direct relation to coding, it might be possible that they can be used to select suitable dictionary units because both represent information about the relative contribution of points in the time-frequency plan. As such they might short circuit the demanding search process associated with matching pursuit.

Zivanovic et al. (Zivanovic et al., 2008) introduced a method of classifying peaks in the DFT-domain, which is similar to the approach presented here, i.e. they integrate over the bins belonging to a peak. Their features to classify a peak as noise, tone or side-lobe resemble an area, not unlike our selection criterion. The advantage of our method is the absence of the need to search for bins belonging to a peak.

Amplitude and frequency modulation will result in cutting up the above-threshold region into the regions of the individual beats for sufficiently high levels of noise (depending on the depth of the modulation). If the size of these area's becomes smaller than our threshold size for noise, the signal will be discarded if no knowledge-driven algorithm searches for regularly repeating beats.

The result of the TF filter could be used as a bootstrap for many other techniques that require a reasonable estimate of the number of tones in a signal. Examples are sinusoidal modeling (Marchand and Depalle, 2008; Röbel, 2007), where the search would only have to be started in a few regions, and gaussian-mixture-model based methods (Roux et al., 2007) where the number of gaussians in the mixture needs to be known.

3.4.3 Relation with human processing

The relation between the local SNR and the fraction of the signal component found in figure 3.5 is intriguing because it lies close to the lower range of the values of the local SNR reported by Fletcher (reviewed in (Allen, 1994)). Although Fletcher estimated a range of 0 to 30 dB local SNR and the range here is about 0 to 20 dB, they can be related. In the case of Fletcher the range corresponded to improvements in vowel recognition probabilities and not to fractions of signal component found, which should be considered as only a small part of the process of phoneme recognition.

This is the first report of a low level signal description that directly relates the local SNR to probabilities. Moreover, the measure is also defined on a similar range as Fletcher. This suggests that some variant of the tone-fit and the pulse-fit might indeed play a role in the auditory system by translating complex time-domain patterns into level independent information that relates spectro-temporal information to physically and statistically meaningful information. This role might be of central importance in understanding the reliability of the auditory system, since it opens a controlled way to couple signal-driven and knowledge-driven processing with the optimal use of physical information.

3.5 Conclusion

Together, the experiments have shown that the tone- and pulse-fit form an efficient, robust, and informative low level signal representations that comply with the requirements listed in section 3.1.3. The experiments have shown that TF and PF are able to indicate all tones and pulses with a high local SNR (figures 3.3, 3.5, and 3.6). Therefore they are, unlike dynamically estimated adaptive background models, insensitive to threshold estimations errors (figure 3.3, lower panel). The measure is dependent on the local SNR

in a predictable and by and large channel-independent way (figure 3.5), expressed in standard deviations of white (or more generally broadband) noise (figure 3.6). Moreover the measure has been shown to be suitable for chirps as an intermediate signal class between sinusoids and pulses. Slow sweeps are well described by the tone-fit, very steep ones by the pulse-fit and intermediate values by a superposition (figure 3.4). This set of desirable properties provides information that can be used by a later signal component tracking stage. All in all, the tone- and the pulse-fit provide time-frequency regions where reliable tonal and pulsal evidence can be derived about the sources that contribute to the signal. These regions have properties that are suitable for environmental sound recognition and results on environmental sound recognition using the tone- and the pulse-fit measure can be found in chapter 4.

4

Sound event identification through expectancy-based evaluation of signal-driven hypotheses

This chapter first appeared as: Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C. Andringa. Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters*, 2010. Work on the dynamic network model as described in section 4.2 is by M.E. Niessen.

We present the results of an experiment where signal-driven (bottom-up) recognition is combined with knowledge of the context (top-down knowledge) to improve the performance of environmental sound recognition in real-world circumstances. The real-world sonic environment is often referred to as a soundscape, that is, an environment of sounds with emphasis on the way it is perceived and understood by an individual or by a society (Schafer, 1977). Although full soundscape analysis is beyond the scope of this chapter, we aim to build a system that can become the basis for an automatic soundscape analysis tool by identifying sound events in real-world environments.

A system that identifies sound events in continuous recordings has ad-

ditional requirements compared to a system that classifies sound samples, of which is known that they have content. In recognition, a system needs to segment the signal and separate the sources before it can classify them (Shinn-Cunningham, 2008; Griffiths and Warren, 2004; Roman et al., 2006; Barker et al., 2005).

Furthermore, a system that analyzes soundscapes has to deal with transmission effects such as concurrent sources and reverberation. Reverberation results in a mixing of the target sound with a time delayed version of itself. Therefore, it precludes the successful application of feature vectors that describe the whole spectrum, such as Mel-frequency cepstral coefficients (MFCC's) and the continuous wavelet transform (CWT). MFCC's have been shown to be very successful for single-source, non-reverberant speech recognition (O'Shaughnessy, 2008). Moreover, MFCC's and CWT have been used successfully in environmental sound recognition provided that the recordings contain a single, clean source (Cowling and Sitte, 2003). However, this is an unrealistic approximation for actual environmental sounds.

Real-world environments pose another problem on techniques used in speech recognition. Speech recognition relies on a strong temporal ordering, but for environmental sounds this ordering is far weaker. Speech recognition techniques exploit this ordering by applying hidden Markov models to find the best model sequence (O'Shaughnessy, 2008). In the case of non-speech sound recognition, such as music genre determination, it has been shown that temporal information is not necessary to recognize genre (Aucouturier et al., 2007). However, music genre determination does not require the detection of sound events and is therefore not suitable to describe the sonic environment in detail.

Another method for sound analysis, the bag-of-frames (BOF) method, has been shown to be able to identify scenes from real-world recordings (Aucouturier et al., 2007). However, the BOF method is not designed to represent details about individual sources in the signal, because it uses long-term statistics of the complete spectral range. Nevertheless, information derived with BOF methods may provide contextual information to guide the classification of sound events.

In contrast to the BOF method and whole spectrum descriptors, the methods we present in this chapter segments the spectrum on the basis of the local spectro-temporal properties. Segments are likely to stem from a single source when they are based on local properties. The robustness and reliability of these segments, called signal components, are improved with grouping principles from auditory scene analysis, such as common onset, common offset and common frequency development (Bregman, 1990; Ellis, 1999). These groups are classified as sound events using a naive Bayes classifier.

Systems that perform environmental sound recognition, with similar preprocessing as proposed in this chapter, are applied commercially in real-life situations (van Hengel and Andringa, 2007). These systems extract one bit

of information from their environment, namely: “is there verbal aggression, or not?”. The more general problem of environmental sound recognition is more complex, but shares some properties with information retrieval, especially with associative retrieval (Crestani, 1997). For both applications it is desirable to retrieve relevant information that is associated with some information item, such as a user query. In environmental sound recognition, retrieval corresponds to estimating the presence of sources and processes from the signal’s history and its environmental context. Similar to information retrieval, it is not essential to recognize all sound sources (documents). Instead, it is important to determine sufficient information about the environment to extract relevant parts of the signal, that is, being able to answer the question that spawned the search. Because of the similarities between environmental sound recognition and associative information retrieval, we use the same measures of success, such as precision, recall, and the F-measure.

The dataset used in this chapter is created to test aggression detection systems. However, the content is fairly rich, since it is recorded on a busy train station. Therefore, it includes problems of real-world environments, such as transmission effects and ambiguous sound events. For example, the sound of a train and a subway are very similar. Based on the sound alone, even human listeners have problems identifying the event correctly, unless they are provided with context (Ballas and Howard, 1987). An automatic system that identifies sound events in real-world situations can benefit from contextual information to recognize events, similar to humans listeners.

To approach this human strategy, we propose a method inspired by cognitive research (Quillian, 1968; McClelland and Rumelhart, 1981). This method constructs a dynamic network that keeps track of both bottom-up signal information and contextual knowledge. By using more information than what can be known from the signal at each point in time, the system is not only more robust to noise, but it can also distinguish between sound events that are similar in acoustic structure but different in meaning (Niessen et al., 2009b). The nodes of the dynamic network represent information about sound events at different levels of complexity. Whenever new signal-driven information becomes available, the information in the network is updated. Subsequently, this information is used to form expectancies of future sound events.

This chapter is divided in five sections. The following section discusses the dataset. Furthermore, we explain the signal-driven processing signal components and machine learning. The third section describes how contextual knowledge is learned and incorporated in the system. Section 4.3 discusses the results of the signal-driven and the combined system, which uses knowledge of the context on top of the signal-driven information. Finally, in the fifth section we explain and discuss the results and give suggestions for future work.

Table 4.1: The annotated classes and the number of their occurrences in the dataset.

class	#
singing	82
speech	521
train	15
subwayDoorSignal	14
subway	40
kick	26
scream	290

4.1 Signal-driven processing

4.1.1 Dataset

The dataset (chapter 6, Zajdel et al., 2007) consists of 40 enacted scenes from 16 different scenarios, which last between 1 and 2 minutes each. The total duration of the recordings is 54 minutes. The scenes were acted by professional actors (three men, one woman) on a platform of the station Amsterdam Amstel. The recordings are distorted by reverberation, because the Amstel station is a glass box. The platform was in normal use by trains on one side and subway trains on the other side. The actors took turns in playing the scenes. For example, the "pickpocket" scenario was played out twice with different actors. All scenarios were played out twice or more. The 16 scenarios were based on stories occurring at stations, such as friends meeting, enthusiastic football supporters and diverse forms of verbal aggression and vandalism. The scenes were recorded by 8 microphones (16 bits, 44.1 kHz sampling rate), of which one was used for this study. This microphone was located about 2 meters from the centre of the action and about two meters from the subway track. Saturation of the microphones was checked not to occur when goods trains passed. The scenes were also captured by three calibrated cameras.

The 40 scenes were annotated by the authors for 7 classes (see table 4.1), based on audio and video. The start and stop times of each event were annotated. For subways and trains, and for some speech, singing and screams, these times were ambiguous, because it is hard to indicate the exact time these events become loud enough to be detectable. The assignment of classes included subjective decisions like whether or not a sound is speech or a scream. These decisions were left to the annotator. Therefore, the annotations are far from perfect (see chapter 5).

4.1.2 Signal components

All recordings are processed using the methods from chapter 3. The cochleograms are created using a 100 channel gamma-tone filterbank and using a frame-

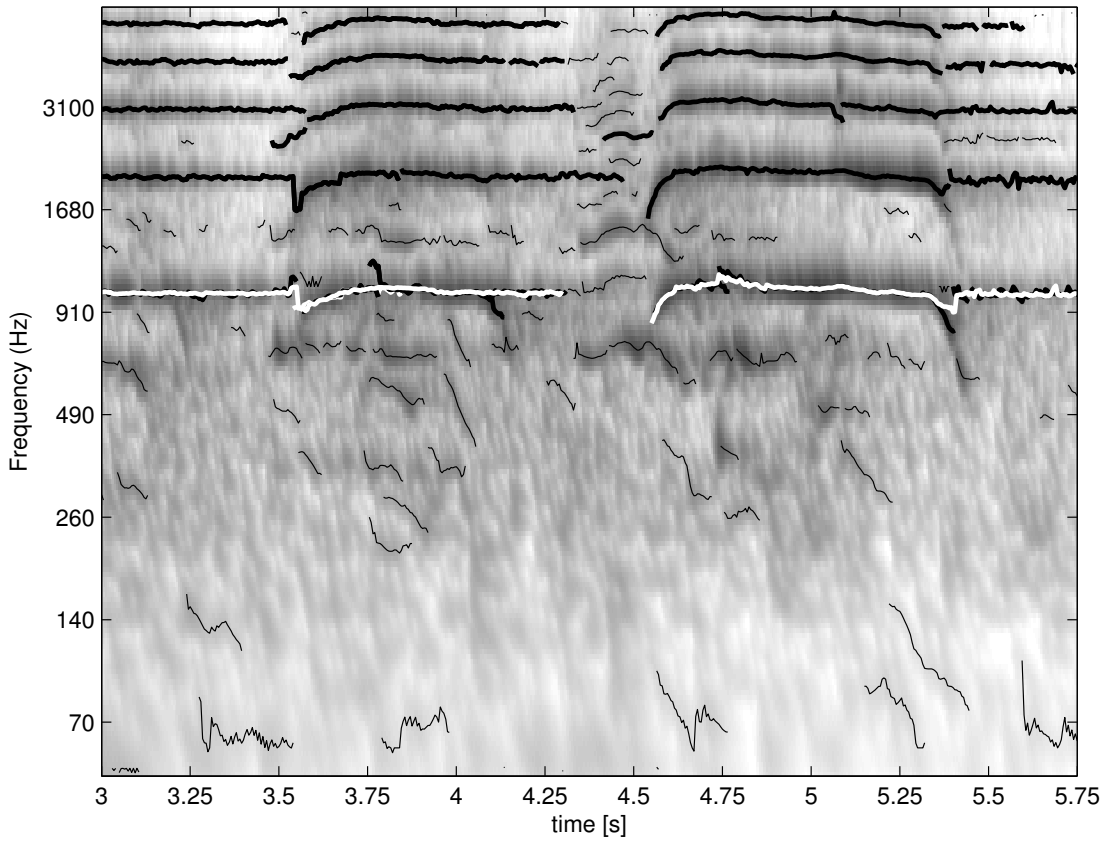

3100
1680
910
490
260
140
70
Frequency (Hz)
3
3.25
3.5
3.75
4
4.25
4.5
4.75
5
5.25
5.5
5.75
time [s]

Figure 4.1

size of 5 ms. The tone-fit and pulse-fit representations are thresholded to create a binary mask. This threshold is set to twice the standard deviation of the TF or PF when applied to white noise. Areas that are too small to be either valid tones or valid pulses are discarded. This pruning, in combination with the mask threshold, limits the number of spurious areas that are caused by broadband signals, while allowing tonal or pulse-like signals. Within the remaining areas the energy maxima of the cochleogram are strung together horizontally to form tonal, or vertically to form pulse-like signal components (see figure 4.1).

4.1.3 Harmonic complexes

If possible, the tonal signal components are combined into harmonic complexes (HC) by selecting more and more tonal signal components that comply with the properties of a harmonic complex. Harmonic complex formation starts by selecting concurrent signal components that have a harmonic relation. These hypotheses generate new hypotheses at fundamental frequencies in the range between 300 and 1200 Hertz by shifting harmonic positions of the signal component. These hypotheses are extended with more and more signal components. The process ends by selecting the hypotheses that comply best to a well-formed HC by maximizing score S:

$$S = n_{sc} + b_{f0} + n_h - \sum_{sc} \mathrm{rms}_{sc} - \sum_{sc} \Delta f_{sc} \tag{4.1}$$

where n_{sc} is the number of signal components in the group, b_{f0} is one or zero depending on the existence of a signal component at the fundamental frequency, n_h is the number of sequential harmonics in the group, rms_{sc} are the root mean square values of the difference of a signal component and the fundamental frequency after the mean frequency difference is removed, and Δf_{sc} is the mean difference between the fundamental frequency and the frequency of the signal component divided by its harmonic number.

For each harmonic complex we calculate nine features, listed in table 4.2. These features will be used in the signal-driven recognition stage.

4.1.4 Broadband events

Evidence for broadband events, such as trains, is determined by an algorithm that searches for slow broadband changes in the signal. These events have to satisfy a combination of criteria. The change in signal must last at least 2 seconds, and 30% of the frequency channels must be more than 6 dB above the long-term background. The long-term background is calculated per channel as the energy value that is exceeded more than 95% of the time. This level of 95% assumes that each channel is dominated by background

Table 4.2: The features extracted from each harmonic complex. Features 1, 2, 3 and 9 are picked by the authors to indicate the strength of the harmonic complex. The other features are selected from van Hengel and Andringa (2007); Zajdel et al. (2007) to discriminate speech, scream, singing and subwayDoorSignal.

#	description
1	length in seconds
2	score from eqn. 4.1
3	number of signal components
4	mean energy under the signal components
5	std deviation of energy under the signal
6	spectral tilt of the signal components
7	mean fundamental frequency
8	standard deviation of fundamental frequency
9	feature 2 divided by feature 1

noise at least 5% of the time. The criterion is fairly safe and works well in practice, assuming that a temporal scope can be chosen appropriately. We chose a temporal scope that was as long as the whole file (about a minute). The energy must exceed the background by three standard deviations of white noise in that channel.

The events that comply to the aforementioned criteria are described with a feature vector of 20 features. The first 15 features are three properties calculated in five frequency bands. Every frequency band contains 20 channels. The 5 remaining features are the first five cepstral coefficients that describe the spectral envelope. The three properties for the five bands are only computed for the 10% most energetic time-frames per event. The first property is the correlation between points in time separated by half a second. This correlation is typically high for slowly changing events and low for fast changing events, such as speech. The second property is the distance between the frequency band and the average energy, in terms of standard deviations of white noise. This property is level-independent and reflects the energy distribution over the bands. The distribution can be different between subway trains and normal trains. The third property is the average foreground-to-background ratio for each band, which reflects the total energy per band compared to the background. This property might differentiate between nearby and far-away events.

4.2 Dynamic network model

The signal-driven processing provides hypotheses based on information in the signal. However, real-world sound recordings, such as in the dataset used in this study (see section 4.1.1), are distorted by transmission effects similar to broadband noise. Furthermore, some sound events can produce similar acoustic signals, but have a different meaning. For example, although speech

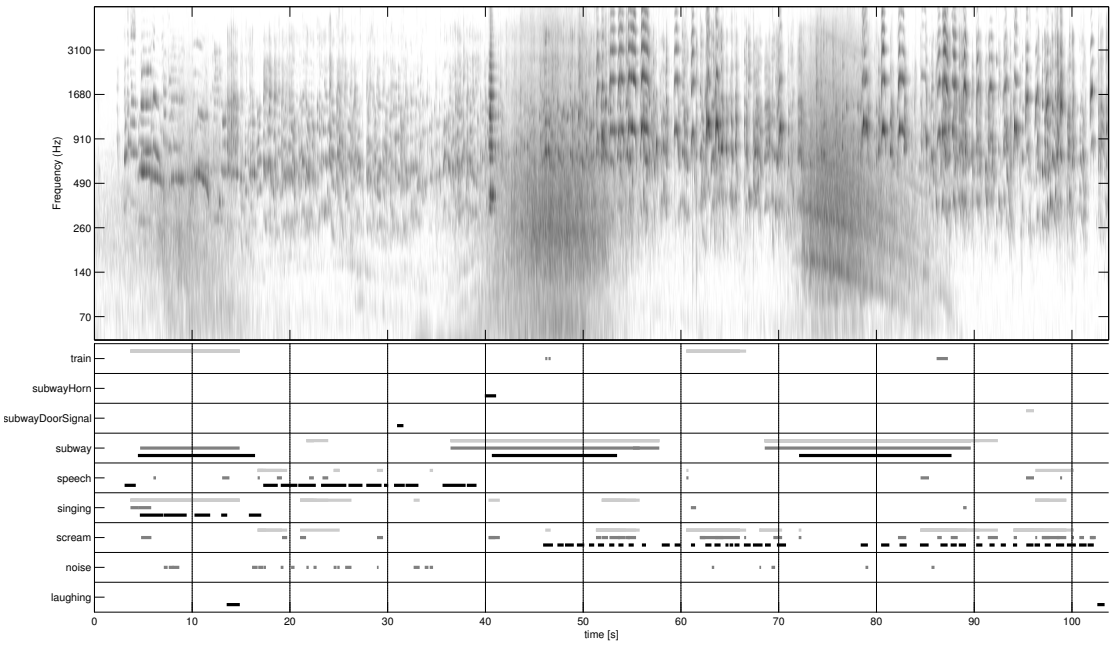


Figure 4.2: The background shows the cochleogram of a few screams and a departing subway train. The black lines indicate signal components, the thick black lines are grouped together to form harmonic complexes, and the white lines indicate their fundamental frequencies. Spurious contributions, due to pattern in noise are inevitable, but they can be discarded if they do not contribute to patterns at higher levels of aggregation.

and screams result in a similar acoustic pattern, they differ in meaning, and require a different response. Distortions due to transmission effects and ambiguous sounds might lead to erroneous hypotheses, because the signal provides too little information to allow a correct inference. Knowledge about the environment and the context of a sound event can be used to improve the classification through predictions. Specifically, past sound events can lead to expectancies of the sound events that will follow. If a signal-driven hypothesis matches an expectancy, it is more likely to be correct. In this section we present a model that creates expectancies of sound events and evaluates the signal-driven hypotheses based on these expectancies. The description of the way the model operates is given in more detail in (Niessen et al., 2009b).

4.2.1 Knowledge network

The knowledge about the environment is learned in a supervised training phase and stored in a static network, referred to as the knowledge network. This knowledge network is similar to semantic networks used in information retrieval (e.g. Crestani, 1997; Maanen et al., 2008). Information retrieval is concerned with retrieving relevant information associated with some information item, such as a user query. Therefore, semantic relations, like similarity, between pieces of information are stored in a semantic network. Nodes in this network represent information items, and the connections between the nodes represent the relations between these pieces of information. In automatic sound recognition, a node could represent a speech event, or a whistle followed by a train arrival. Furthermore, the relation between events are represented by the strength of their connection.

Annotations of sound recordings (see section 4.1.1) are used in the supervised training phase to learn relations between sound events. When two sound events occur within a certain interval, they are combined in a separate node. The relation between the node that represents the sequence of the events and the nodes that represent the individual sound events is calculated according to a term-weighting approach used in automatic document retrieval (Salton and Buckley, 1988). In this method the importance of a term (word or phrase) in a document is determined by multiplying its frequency in the document with the inverse frequency it occurs in other documents. Hence, the term is important for a document if it occurs often in that document and infrequently in other documents. Analogously, if a sound event A is encountered often in combination with some sound event B, and little with other sound events, it is important in the event sequence $S : A-B$. Accordingly, the strength between the sound event A and the event sequence S is:

$$w_{A,S} = \mathrm{tf} \cdot \log\left(\frac{N}{n}\right), \tag{4.2}$$

where N is the total number of sequences, n is the number of sequences in which A occurs, and the term frequency is given by:

$$\mathrm{tf} = \frac{f_{A,S}}{\sqrt{f_A}}, \tag{4.3}$$

where $f_{A,S}$ is the number of occurrences of A in S, and f_A is the total number of occurrences of A in the training set.

Most sequences represent events that can occur in any order. For example, sound events produced by people, such as singing and speech, will generally be heard together, but not in a fixed order. However, for some sequences the order can be very indicative. For instance, in the dataset there are trains departing, which are always preceded by a whistle of the conductor. Hence, if a whistle is heard, a strong expectancy of a train departing should arise. To capture the expectancies of fixed sequences, we determine whether the sound events that constitute a sequence have a strong bias to a specific order. For these fixed sequences the mean time difference between the events is used in a function to calculate the expected value of the second event in the sequence. In other words, the first sound event of a fixed sequence primes the network for the second sound event after a learned time interval. In the next subsection we will show how this expected value is computed for both ordered and non-ordered sequences.

4.2.2 Dynamic network of hypotheses

Once the knowledge network is fully trained, it is used in the operation phase to evaluate signal-driven hypotheses of sound events. Each signal-driven hypothesis is initiated as a node in the dynamic network. The dynamic network has three levels of representation. The hypotheses at the first level represent detected structures in the signal, as described in section 4.1. The second level consists of hypotheses of possible sound events that explain the structures. Finally, the third level contains hypotheses of sequences of events, as described in the previous subsection. Figure 4.3 shows an example of a network with two signal-driven hypotheses about structures in the signal, their connections to possible sound events that caused them, and a sequence of which they might be part.

When a new signal-driven hypothesis is added to the dynamic network, the configuration of the network is updated. First, the hypothesis that represents a structure in the signal is connected to hypotheses of sound events that can explain the structure. The strength of this connection is determined through naive Bayes classification of the structures, as will be described in section 4.3.1. Next, the connections of these sound events to possible event sequences are retrieved from the knowledge network and added to the dynamic network. The connections in the network are only between hypotheses

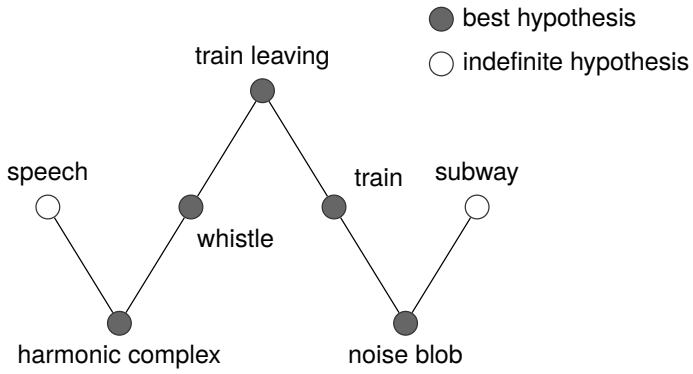


Figure 4.3: An example of a network with two signal-driven hypotheses about structures in the signal. Both hypotheses are connected to two hypotheses of sound events that can explain the structure. Two of these hypotheses are part of an event sequence, increasing the support for the sound events that are part of the sequence.

at different levels, as can be seen in Figure 4.3. As a consequence, the dynamics and hierarchy of the network are captured by the hypotheses and their connections.

4.2.3 Activation

The activation value of a hypothesis is a weighted sum of its input activation from connected hypotheses. The activation of a signal-driven hypothesis is spread through the network after the configuration is updated. As a result, every hypothesis in the network holds a confidence value after spreading the activation. A description of the details of the spreading activation algorithm can be found in (Niessen et al., 2009b). The activation values of all hypotheses in the network decrease with time when they get no reinforcement from signal-driven evidence.

The activation values of event sequences are used to compute the expected activation of events that are not active yet, and are part of the sequence. For example, in a non-fixed event sequence such as singing and speech, of which speech is already identified, the expected activation of a singing event is calculated by multiplying the activation value of the event sequence with the connection strength between the sequence and the type of event (see Formula (4.2)). Since the activation value decays with time, the expected value is smaller when the other event of the sequence occurred longer ago.

For fixed event sequences, the expected value will furthermore be dependent on the time when the event is expected:

$$\hat{A}_i(t) = w_{ij} A_j(t - \Delta t) \mathrm{e}^{\frac{-(\Delta t - \bar{T})^2}{2\sigma^2}}, \tag{4.4}$$

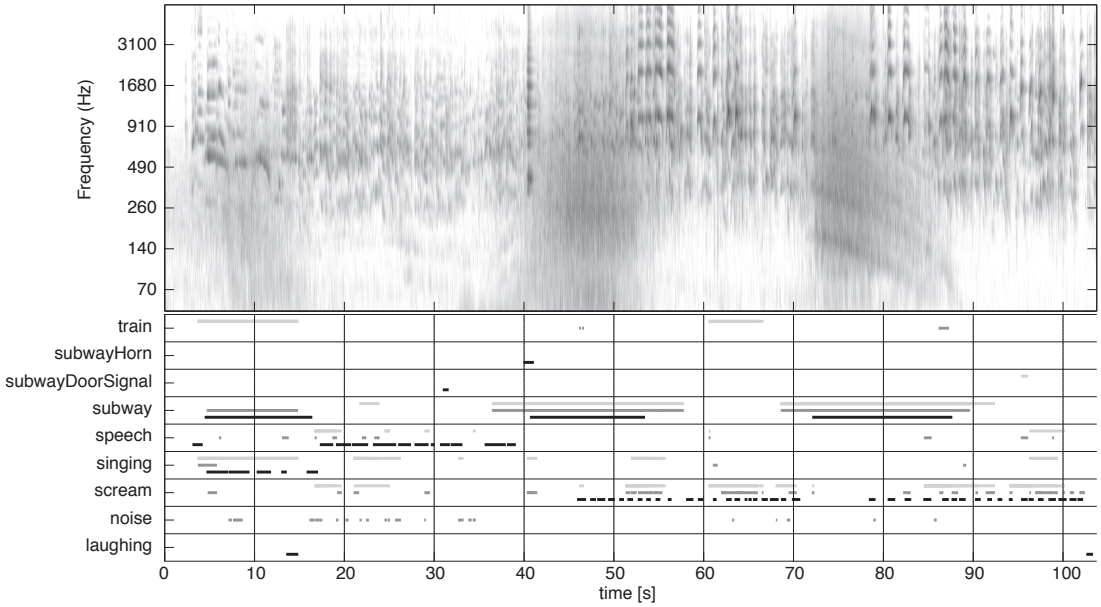


Figure 4.4: The upper panel shows the cochleogram of a complete scenario. A darker color corresponds to more energy. In the first 40 seconds there is some speech and singing. At t = 41 s a subway horn occurs, which is followed by the noise event of a subway train passing by. Around t = 55 s four clear screams occur, followed by a few more muffled ones. At t = 72 s a subway train enters the station, again followed by screams. The lower panel shows the annotations and detections for the different classes. The lower, black, lines represent the annotations, the middle, gray, lines the signal-driven detections, and the upper, light-gray, lines the final, expectancy-based results.

where w_{ij} is the connection strength between expected sound event i and event sequence j, $A_j(t-\Delta t)$ is the previous activation value of event sequence j, Δt is the time span since j started, and average time span $\bar{T}$ and standard deviation σ describe the time distribution of the event sequence, as it is learned during the supervised training phase.

4.3 Experiments

To test the system we apply it to the dataset of 40 realistic recordings (see section 4.1.1). In the first experiment only the signal-driven classification is used. In the second experiment these results are used in the expectancy-based dynamic network.

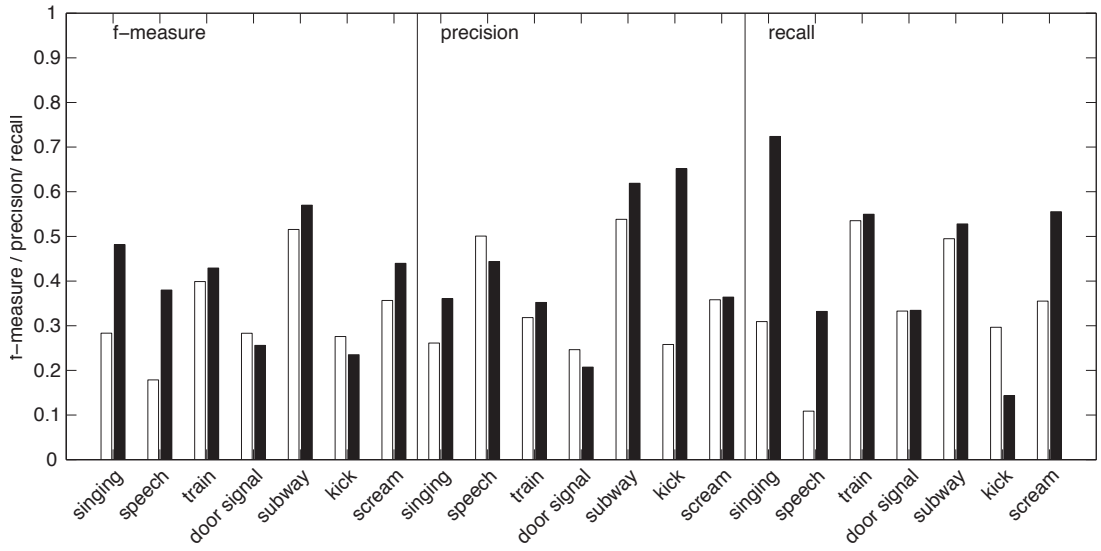


Figure 4.5: The results of the signal-driven classification (white bars) and of the expectancy-based results (black bars).

4.3.1 Experimental setup

All 40 audio files were processed with the methods explained in section 4.1 to extract harmonic complexes and their features (see table 4.2). The harmonic complex with the highest score and overlap was selected for each annotation and labeled according to the annotation. Harmonic complexes that do not overlap in time with an annotation were labeled as noise. Harmonic complexes that do overlap with an annotation, but do not have the highest score, are discarded. From these files, 40 pair files were generated, of which 40 files were used for training, all with the instances from one scene left out, and 40 files were used for testing, with instances from the scene that was left out, thus creating a leave-one-scene-out set.

Because of the strong link with information retrieval (see section 4.2.1) we use performance measures from that field, such as precision and recall, to quantify the performance of our system. Precision is a measure for the fraction of time our detections were correct, and recall is a measure for the fraction of detections we should have made are actually made. The F-measure is the harmonic mean of these two, giving a single performance measure. The formula's are given as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{4.5}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.6}$$

$$F = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.7}$$

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

For the first experiment a naive Bayes classifier from the Weka toolbox ((Witten and Frank, 2005)) was trained on the leave-one-scene-out training file and tested on the corresponding testing file. The labeling and classification of the noise regions was performed in the same way as the harmonic complex classification. The results of both classifications were taken together to create a single result set.

In the second experiment the supervised training of the knowledge network (see section 4.2.1) was performed on the same data as the classifier, that is, the annotations of the leave-one-scene-out training file. Hence, the test set was not used for training. On average 18 different types of sequences were encountered in the training set. These sequences are composed of the 7 classes listed in table 4.1. An average of 89 examples of each sequence was used to train the weights in the knowledge network. The spread of the number of examples per sequence is very large, ranging from 2 to 730. For the testing, the results of the classifier were input for the dynamic network of hypotheses (see section 4.2.2).

4.3.2 Signal-driven results

The white bars of figure 4.5 show the F-measure, the precision, and the recall of the signal-driven classification. The overall F-measure is 0.37, the overall precision is 0.39, and the overall recall 0.34. The results of one of the scenes are shown in the lower panel of figure 4.4.

Part of the errors arise from alignment errors of the annotations. For example, all detections of the subway trains are longer than the annotations. This problem is hard to solve, because the annotators did not agree when on the moment when a train is first and last detectable. Therefore, the detection cannot agree with both annotators. A partial solution would be to introduce "don't care" regions around annotations where the algorithm is not punished for incorrect detections.

The major groups of confusion are between trains and subway trains, and between speech, singing, and screams. These confusions may partially be caused by confusion in the annotations. The distinction between a train and a subway train is hard to make based on audio recordings, even for a human annotator. The boundaries between the classes speech, singing and scream are fairly arbitrary, which causes confusion in the annotations.

The F-measure on the kick class is small because it is neither a harmonic nor a broadband sound. The features we have used were not suited for describing these pulse-like sounds.

The systems calculations run at about real-time on a modern PC (2 GHz dual-core). However, the current Matlab code is not optimized. Based on similar systems optimized for speed (van Hengel and Andringa, 2007) we

estimate that the performance could be around four times real-time on the same machine.

4.3.3 Expectancy-based results

The black bars of figure 4.5 show the performance measures for the classification including the dynamic network. The overall F-measure improves to 0.45 (20%), the overall precision to 0.42 (8%) and the overall recall to 0.49 (44%). The main improvement is in the recall of the classes that have more harmonic content (singing, screams and speech), because events of these classes are more likely to be of the same class as their neighbors. As a consequence, the network may change a speech classification to a scream when surrounded by screams. If this change is correct, both the recall of the scream class and the precision of the speech class increase. However, the increase in precision is moderated by other erroneous changes. As a result, the overall precision does not increase substantially. Due to the ambiguous nature of some of the classes and the acoustic environment a high F-measure is not achieved. So we conclude that the inclusion of the dynamic network leads to a result more consistent with manual annotation.

4.4 Discussion

In the previous section we have demonstrated that the combination of signal-driven algorithms and a dynamic network of hypotheses results in a recognition improvement for most sound event classes compared to an exclusively signal-driven method. Especially the classes that have similar signal structures, and hence rely more on context for their interpretation (screams, speech and singing), are better identified in the combined approach. Classes that are already identified well by the signal-driven algorithm (subway and train) gain little improvement from the dynamic network. Finally, both classes that occur infrequently, and hence have little training examples, and classes that are not yet captured well by the signal features, show a small performance reduction.

We have shown that the use of a dynamic network model improves the overall performance of environmental sound recognition. However, apart from sound event recognition, this model provides more divers ways to analyze a soundscape. More specifically, through hierarchical relations in the network, recognition of sound events can lead to abstract descriptions of the soundscape. This introduces the possibility to describe complex activities in the neighborhood of the microphone with complex and efficient linguistic descriptions (Guastavino, 2007).

Furthermore, the input information that is presented to the network is not limited to a specific modality. In Niessen et al. (2009a) we show that the

dynamic network model can also be used to improve visual robot localization. Because the model can receive input from different modalities, it can combine multiple modalities in a single system. For example, if input from one modality, such as images, is insufficient, input from other modalities, such as audio or GPS, can help to generate predictions. In future work we plan to integrate information from multiple sources of knowledge to reach more reliable event recognition with richer descriptions.

One of the major problems in the development of environmental sound recognition systems that operate in real-life situations is the lack of large, diverse, and annotated datasets that can be used for training and testing. This is one of the reasons that we tested on a dataset that represented only a single location and a limited amount of events. The main problem of constructing more realistic datasets is the large number of different events that can occur outdoors and the associated time it takes to annotate a representative set. The development of an annotation tool for soundscape research is helpful in this respect.

Another problem in environmental sound recognition is performance evaluation. We have used the measures precision and recall to quantify the performance, since these measures are common in the related task of information retrieval. We calculated these measures in terms of the temporal overlap of annotations and classifications. However, if we were to apply these measures in line with the field they were originally developed for, we should only check whether or not an annotated event was detected. We have chosen for overlap instead of presence, because the combination of the short annotations of speech events in combination with small temporal alignment errors made the attribution difficult. Allowing some flexibility in matching system detections with hand annotations may alleviate this problem. This however requires a more formal justification, before it can be applied.

The current system shows that it is possible to build a recognition system that captures many of the events of a realistic and minimally constrained sonic environment. The background was completely uncontrolled while the foreground consisted of actors who improvised a range of both social and aggressive activities. We have shown that it is beneficial to use the history of identified sound events to form a context in which the current sonic evidence is weighted. This is done by forming a dynamic network that mimics short-term memory dynamics. The interplay of knowledge-driven and signal-driven processing is characteristic for human perception. Since human perception is effectual in a wide range of acoustic environments, we consider this interplay a promising approach for robust automatic sound recognition.

5

How to evaluate the sources in a soundscape?

This chapter first appeared as: Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C Andringa. Annotating soundscapes. In *Internoise 2009*, 2009.

5.1 introduction

Humans can recognize events in the sonic environment (soundscape) seemingly effortlessly. However, this ability thus far eludes our technical abilities (Cano, 2006). Automatic sound recognition has important applications in fields as diverse as environmental noise monitoring, robotics, security systems, content-based indexing of multi-media files, and even modern human-system interfaces. Most sound recognition research is aimed at improving one aspect of these application domains, such as speech recognition or music genre detection. These limited domain solutions can rely on domain dependent assumptions that simplify the problem considerably. For example, within music classification (Aucouturier et al., 2007) or speech recognition (O'Shaughnessy, 2008) it is typically assumed that the input does not contain multiple uncorrelated streams of sonic evidence. As a consequence, stream segregation and other problems are defined out of the problem-space and are not addressed scientifically.

In contrast to domain specific solutions, a general sound recognition system should be robust to the complexities of unconstraint soundscapes, such as strong and varying transmission effects and concurrent sources. To handle real-world complexities, human perception relies on signal-driven processing, but also on contextual knowledge and reasoning (Niessen et al., 2009b). Therefore, a general sound recognition system should comprise an interaction of signal-driven techniques and interpretation of the context.

This paper focuses on the development of a tool to facilitate real-world sound annotation for training and benchmark purposes. It uses a set of simple algorithms to detect sonic events and to classify these events. The interaction between semantic content, in the form of annotations, and signal-based evidence forms the basis of future, more general, sound recognition systems. The annotation of everyday sounds must lead to an adequate description of the content of a sound-file in terms of the interval in which an event occurred. Annotation is a time-consuming, and knowledge intensive task, which is usually quite boring as well. This is probably the reason why there is currently only a single annotated database of sounds in realistic everyday conditions (van Grootel et al., 2009). Carefully selected everyday sounds in benign conditions have been used in other studies (Gygi et al., 2007; Marcell et al., 2000). However for these sounds the annotation problem is trivialized, because the datasets contain single sound events in a single file.

There are many difficulties associated with real-world sound annotation: The great within class diversity of sounds (e.g. cars at different distances and speeds) in combination with the co-occurrence of other classes makes it difficult to interpret a visual rendering of the signal as spectrogram and to annotate the visual representation without listening to the sounds in context. Visual inspection of spectro-temporal representations is an important aid for annotation, but attentive listening to the sound is essential. Sonic events are often difficult to recognize using sound as the only modality. It is important to annotate the sound during, or soon after, recording. The use of video information can be very helpful whenever the sound sources are clearly visible and easily attributable (which is often not the case). Anecdotic evidence suggests that annotation by someone who was not present when the sound was recorded is much more error-prone and often many sounds cannot be annotated in detail. For example, the difference between cars, truck, busses, and even motorcycles is usually not at all obvious.

The co-occurrence of multiple qualitatively different sonic events and sound producing processes can lead to very complex signals, e.g. coffee-making in a lively kitchen. In these cases it is difficult to track multiple uncorrelated processes and describe each in detail. One might aim to annotate the so-called foreground or, alternatively, the events that attract attention. However, this creates the new problem of determining what attracts attention or what to assign to the foreground. The large number of individually distinguishable events of a similar kind, such as singing birds in a forest,

entails a lot of repetitive work. Realistic environments contain many barely audible events, e.g. distant speakers, which might or might not be included in the annotation. Not including these might unjustly punish a detection system that detects the valid, but un-annotated, events. Conversely, including even the faintest events is both time-consuming and prone to classification errors.

Finally, the determination of the precise moment of the start and end of audible events is subject to similar difficulties as those in the previous point. Especially the detection of the on- or offset of a gradually developing event, like a passing car in a complex environment, is often quite arbitrary. If the measure of success of a recognition system is based on determining the intervals in which events occur, the system is punished for any deviation of this arbitrary choice. The difference between annotators who were present and who were not, suggests that the sonic evidence may often be insufficient (for the human listener). This poses a fundamental problem for each sound-only annotation or recognition system, whether human or machine; a correct recognition result may simply be impossible. Hence, a perfect ground-truth is not a realistic goal for a real-world sound recognition system. Instead, a performance equivalent to human performance when not present during recording is more appropriate.

The current paper focuses on an annotation tool that helps to provide more insight in these problems and helps to alleviate a number of them. It assists a human annotator by reducing the number of repetitive actions by automatically suggesting annotations based on previous annotations. This allows for the human annotator to accept the suggested annotation simply as an instance of the proposed class, instead of having to select it from a (long) list of possible classes. Within the annotation system we try to maximize the probability that the true event class is on top of the list. Initially this list is simply alphabetic. During manual annotation the class list is reordered according to the estimated probability that a certain event is an instance of the most likely classes.

In the next section, we will give an overview of the annotation system. Furthermore, we present the data on which it is tested. In the third section we will give the results of a pilot-experiment on a set of real-world recordings. The paper ends with a short discussion of the annotation process.

5.2 Methods

In this section we first describe the dataset that is used to test the annotation system. This system is based on processing sound in the spectro-temporal domain. Therefore, the sound signal is first pre-processed, which will be explained in section 5.2.2. Subsequently, we describe how the sound is segmented into regions that are likely to include the most energetic spectro-

temporal evidence of the main sources. In section 5.2.3 we show how these regions are described in terms of a feature vector, and how this feature vector is used to classify the regions. The section is concluded with a system overview, which is shown in Figure 1.

5.2.1 Dataset

The dataset was collected under different weather conditions on a number of days in March 2009 in the town of Assen (65,000 inhabitants, in the north of the Netherlands). The recordings were made by six groups of three students as part of a master course on sound recognition. Each group made recordings of three minutes at six different locations: a railway station platform, a pedestrian crossing with traffic lights, a small park-like square, a pedestrian shopping area, the edge of a forest near a cemetery, and a walk between two of the positions. Recordings were made using M-Audio Microtrack-II recorders with the supplied stereo microphone at 48 kHz and 24 bits stereo. This data, with annotations by the students, will be made available on http://daresounds.org.

5.2.2 Preprocessing

The first processing is the transformation of the audio signal to the time-frequency domain using the techniques described in section 2.3. From the resulting cochleogram the tone-fit and pulse-fit values as described in chapter 3 are calculated. The cochleogram is used in the next section for segmentation and the cochleogram, tone-fit and pulse-fit representation are used in the feature vectors as described in section 5.2.3.

Segmentation

The segmentation strategy is fairly basic. It is aimed at the inclusion of spectro-temporal maxima in the form of blobs in the spectral and/or temporal direction. These blobs become prominent by subtracting a strongly smoothed cochleogram from the original. The cochleogram is smoothed in the temporal direction through leaky integration with a time-constant $\tau = 5s$. The time constant τ determines the separation between fast, typically foreground, sonic events and slow, typically background, events. The leaky integration operation corresponds to a delay in the expression of mean energy values that is corrected by time-shifting the resulting values backwards with the time-constant. This time-shift leads to a delay equal to the time-constant, which is not problematic for off-line processing, but that is not desirable for online and real-time processing. The temporal smoothing of time-series $x(t)$ to yield $x_s(t)$ is defined by:

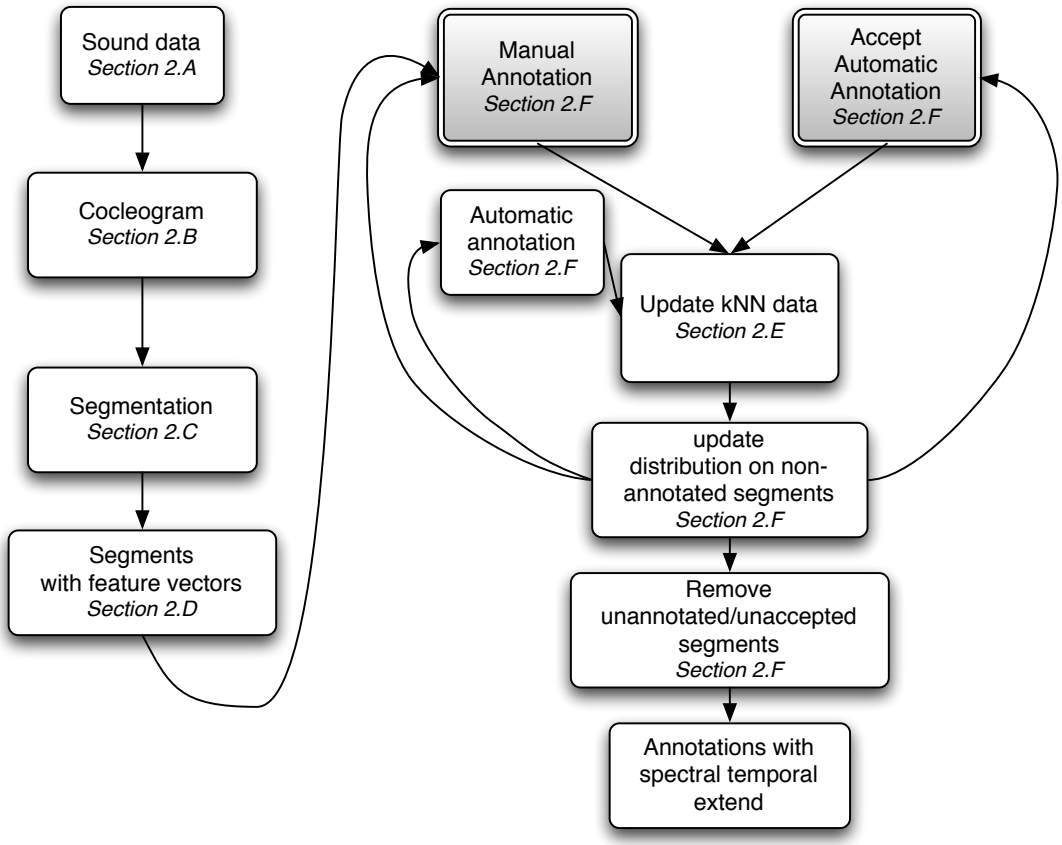


Figure 5.1: Overview of the assisted annotation system. The two gray blocks are the only places of human intervention.

$$x_s(t) = x(t - \delta t)\exp(-\delta t/\tau) + x(t)(1 - \exp(-t/\tau)) \quad (5.1)$$

δt denotes the frame step of 5 ms. In addition to temporal smoothing, the cochleogram is also smoothed in the frequency direction by taking a moving average over 7 channels. The difference between the original cochleogram and the smoothed cochleogram can be termed a fast-to-slow-ratio and is expressed in dB. The regions with a fast-to-slow-ratio of more than 2 dB are assigned a unit value in a binary mask. This mask is smoothed with a moving average in both the temporal direction (25 ms) and the spectral direction (5 channels). The final mask is obtained by selecting average mask values greater than 0.5, which smoothens region perimeters and reduces the number of supra-threshold time-frequency points in the inner-regions of the mask that lead to small holes in the mask. The final segmentation step is the estimation of individual coherent regions in the mask and to assign a unique number to each region. The smallest bounding box that contains the whole region is used to represent the region graphically (see figure 3). There are no special safeguards to ensure either that each region represents information of a single source, or that all information of the source is included in the regions. For example, when two cars pass at approximately the same time, a single region will represent both. Alternatively, sounds that are partially masked by (slowly developing) background sounds tend to break up into a number of smaller regions, that are each less characteristic of the source. Nevertheless, the current settings seem able to include important source information of a wide range of sources.

5.2.3 Feature vectors

The feature vectors must describe the source information represented by the regions. The 37-dimensional feature vector represents properties related to the physics of the source. Note that normal approaches to environmental sound feature estimation (Cowling and Sitte, 2003) make no effort to include source physics other than representing frequency content. The use of the TF/PF-values allows us to attribute signal energy to tonal, pulse-like, or noisy contributions, which result from either source limitations or transmission effects. Table 1 describes the feature vector. The feature vector reflects the channel contributions per region, the fast-to-slow ratio, and the distribution of tonal (TF) and pulse-like (PF) contributions. These signal descriptors are represented by 7 different percentile values from the histogram of the local indicators. Different percentile values might be indicative for different classes. For example, the 90 and 95 percentile values might be highly indicative for footsteps in noise, while the other percentiles might not discriminate from a the noisy contribution in a car passage.

Table 5.1: Region feature vector description

Feature	Dim	Percentile or range	Description
Size	1	> 0.02	Fraction of spectro-temporal area equivalent to 1 s
Channel mean	1	1 - 100	Average channel number (1 is highest, 100 is lowest). This corresponds to average log-frequency contribution.
Channel std	1	< 50	Provides a single number indication of the channel spread.
Fast-to-Slow-Ratio	7	[5 10 25 50 75 90 95]	The distribution of Fast-to-Slow-percentiles provides information about the distribution of strong foreground values
TF	7	[5 10 25 50 75 90 95]	The distribution of TF values provides information about the distribution of strong sinusoidal contributions.
PF	7	[5 10 25 50 75 90 95]	The distribution of PF values provides information about the distribution of strong pulse-like contributions.
Channel distribution	7	[5 10 25 50 75 90 95]	The channel distribution provides more detailed information about the pattern of contributing channels.
Channel spread	3	5-95, 10-90, 25-75	Provides more detailed information about the channel spread as the difference in channel numbers between three percentile pairs of the channel distribution
Frame spread	3	5-95, 10-90, 25-75	Provides more detailed information about the temporal spread as the difference in frame numbers between three percentile pairs of the frame distribution

Classification

Classification of regions based on the feature vector must lead to proposed classes for regions similar to annotated regions. The classifier should function in an on-line fashion and must not require long re-training phases. Additionally, the classifier should be able to function with minimal training data. This combination of demands suggests a simple k-nearest-neighbor (kNN) classifier (Duda et al., 2000). Such a classifier stores all training feature vectors in a matrix. It classifies each region by calculating the Euclidian distance d to all vectors in the training matrix and selecting the k closest training examples which each represent an example of a single class. A simple majority voting system is used to determine the best class for the region. To create a distribution over multiple classes we count the number of occurrences a class in the top k = 5 and divide this by $\frac{kd}{\sum d}$ to get a number indicating the match.

System overview

An overview of the annotation system is given in Figure 1. The system loads, pre-processes, and segments the data of a single file and presents the result to the user. First, the user selects a region. The selected region can be played as sound and a matching class can either be selected from a class-list or added to the class-list. Initially the list is ordered alphabetically, but when sufficiently matching examples of the class have been encountered, the top-positions on the list will be ordered according to class-likelihood. After class assignment, the kNN training matrix is extended with the feature vector of the region. If the match of a class exceeds a threshold (here set to p>0.04), it is automatically classified as that class. If the match exceeds 0.01, the region will be conditionally classified, which entails that the user has to accept the classification before it is included in the kNN training matrix. Regions that end up without annotation are discarded after the user decides that the file is annotated in sufficient detail. To measure the performance of the system we track the class-rank of manually annotated regions, the number of automatically annotated regions, and the number of accepted regions. The number of discarded regions is a measure for the performance of the segmentation. The final output of the system is a list of classes assigned to regions.

5.3 Results and discussion

Measuring the performance of the system in meaningful numbers is difficult. A sensible measure is the time saved by this system compared to full manual annotation of start and stop times of the sound events. However

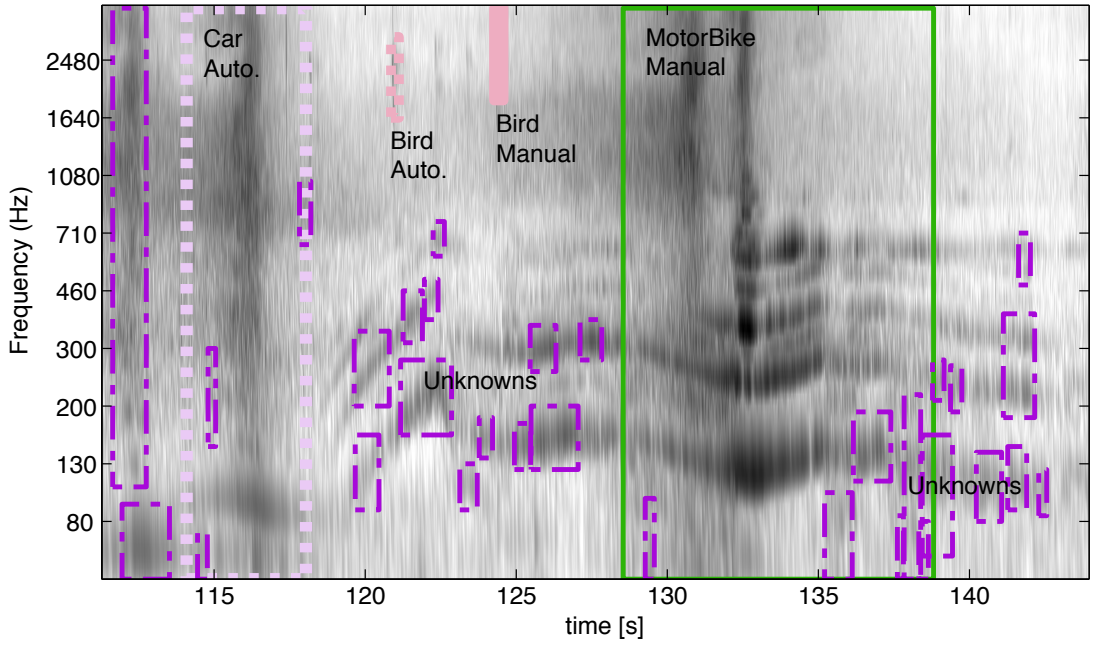


Figure 5.2: The cochleogram of several passing cars. Darker means more energy. Solid lines denote manual annotation. Dashed lines denote automatic classification, the dash-dotted lines denote still unclassified regions. The cars are segmented in the pink boxes. A bird is segmented in the green box. The purple boxes are not (yet) annotated.

Table 5.2: Results of an annotation session on the Assen dataset ($N = 101$)

alphabetical	first	second	total
15%	74%	13%	100%

each annotation session will result in different annotations due to the reasons formulated in the introduction. This makes a fair comparison difficult. Furthermore the current system is not yet sufficiently user-friendly to allow a good comparison. Alternatively we measured how often the correct class was suggested by the kNN classifier. The results are shown in table 2. When a class is either not annotated yet or misclassified, it is marked as "alphabetical", otherwise it is ranked as first or second. Without automated annotations one expects an average rank equal to half the number of classes. Note that with k = 5 it is possible to have 5 different classes in the list, but third, fourth or fifth ranked classes did not occur in the test. The current system is a first installment of the annotation tool. Its initial performance is encouraging, but each aspect can and must be improved before it is truly useful. The further improvement of the tool will depend strongly on an improved understanding of the annotation process, which in turn is a special form of listening. Initial experience with assisted annotation indicates that the annotator does not analyze the file from start to end, but instead prefers to focus either on individual environmental processes or on individual auditory streams. This allows maximal benefit from process/stream dependent knowledge. It is possible that everyday listening (Gaver, 1993b) reflects this so that at most one stream is analyzed with all available knowledge: the focus of auditory attention. All other streams are analyzed in less detail. This observation in combination with and the annotation problems formulated in this paper suggest that the question "What do we do when we listen" should become a focus of active research.

Part III

Evaluation

6

Ambient awareness: Aggression detection

This chapter first appeared as: Wojtek Zajdel, Johannes D. Krijnders, Tjeerd C. Andringa, and Dariu M. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *Proceedings of 2007 IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 200–205, 2007.

Surveillance technology is increasingly fielded to help safeguard public spaces such as train stations, shopping malls, street corners, in view of mounting concerns about public safety. Traditional surveillance systems require human operators who monitor a wall of CCTV screens for specific events that occur rarely. Advanced systems have the potential to automatically filter-out spurious information and present the operator only the security-relevant data. Existing systems have still limited capabilities; they typically perform video-based intrusion detection, and possibly some trajectory analysis, in fairly static environments. In the context of human activity recognition in dynamic environments, we focus on the relatively unexplored problem of aggression detection.

Earlier work involved solely the video domain and considered fairly controlled in-door environments with static background and few (two) persons (Datta et al., 2002). Because events associated with the build-up or enactment of aggression are difficult to detect by a single sensor modality

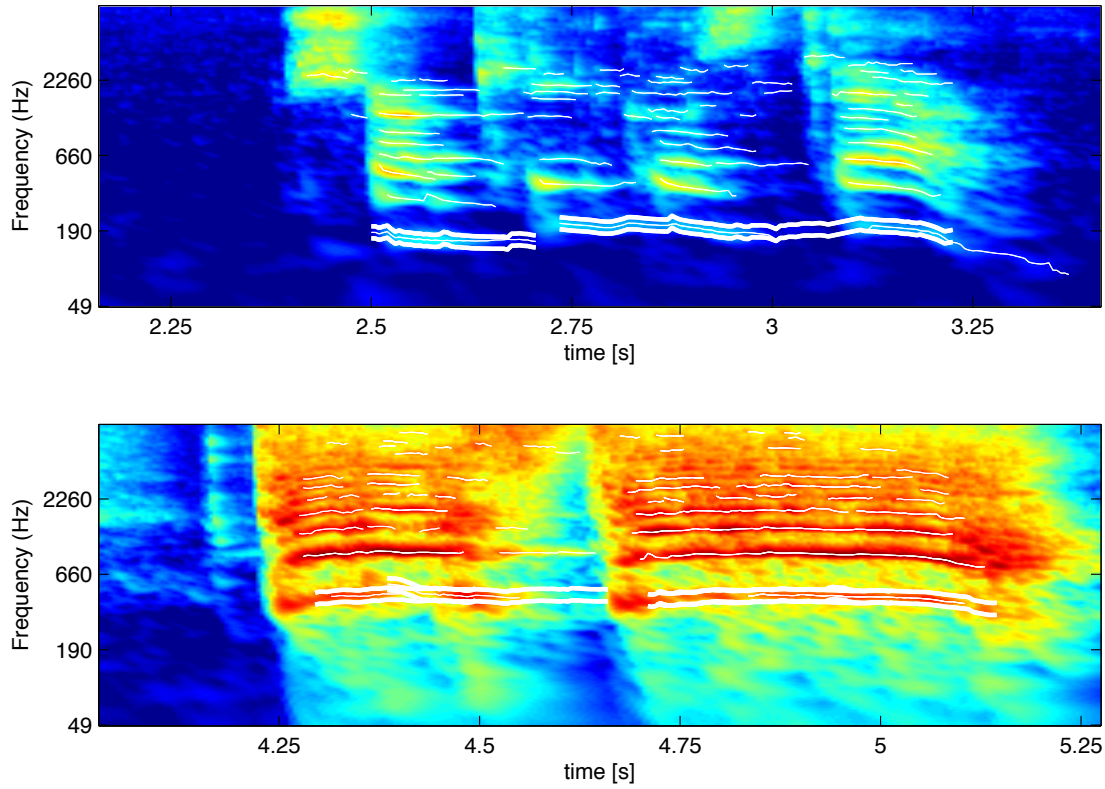


Figure 6.1: Typical cochleograms of aggressive and normal speech (bottom an top figure, respectively). Energy content is color-coded (increasing from blue to red). Note the higher pitch (marked by thick white lines) and the more pronounced higher harmonics for the aggressive speech case.

(e.g. shouting versus hitting-someone), in this work we combine audio- and video-sensing. At the low level raw sensor data are processed to compute "intermediate-level" events or features that summarize activities in the scene. Examples of such descriptors implemented in the current system are "scream" (audio) or "train passing" and "articulation energy" (video). At the top level, a Dynamic Bayesian Network combines the visual and auditory events and incorporates any context-specific knowledge in order to produce an aggregate aggression indication. This is unlike previous work where audio-video fusion dealt with speaker localization for advanced user-interfaces, i.e. using video information to direct a phased-array microphone configuration Pentland (1996).

6.1 System description

6.1.1 Audio unit

Audio processing is performed in the time-frequency domain with an approach common in auditory scene analysis (Wang and Brown, 2006). The transformation of the time-signal to the time-frequency domain is performed by a model[1] of the human ear (Duifhuis et al., 1985). This model is a transmission-line model with its channels tuned according to the oscillatory properties of the basilar membrane. Leaky-integration of the squared membrane displacement results in an energy-spectrum, called a cochleogram (see figure 6.1, section 2.3).

A signal component is defined as a coherent area of the cochleogram that is very likely to stem from a single source. To obtain signal components, the cochleogram is first filtered with a matched filter which encodes the response of the cochlea to a perfect sinusoid of the frequency applicable to that segment. Then the cochleogram is thresholded using a cut-off value based on two times the standard deviation of the energy values. Signal components are obtained as the tracks formed by McAulay-Quatari tracking (McAulay and Quatieri, 1986), applied on this pre-processed version of the cochleogram. This entails stringing the energy maxima of connected components together, over the successive frames.

Co-developing sinusoids with a frequency development equal to an integer multiple of a fundamental frequency (harmonics) are subsequently combined into harmonic complexes. Note that these harmonics can be combined safely because the probability is small that uncorrelated sound sources show this measure of correlation by chance.

Little or no literature exists on the influence of aggression on the properties of speech. However the Component Process Model from Scherer (Scherer, 1986) and similarities with the Lombard reflex (Junqua, 1993) suggest a couple of important cues for aggression. The component process theory assumes that anger and panic, emotions strongly related to aggression, are seen as a ergo-tropic arousal. This form of arousal is accompanied by an increase in heart frequency, blood pressure, transpiration and associated hormonal activity. The predictions given by the model show many similarities with the Lombard reflex. The increased tension on the vocal chords increases the pitch and enhances the higher harmonics, which leads to an increase in spectral tilt. These properties, pitch (fundamental frequency (f0)) and spectral tilt (a measure of the slope of the average energy distribution, calculated as the the energy of the harmonics above 500 Hz divided by the energy of the harmonics below 500 Hz) are calculated from the harmonic complexes.

[1] We thank Sound Intelligence (http://www.soundintel.com) for contributing this model and cooperation on the aggression detection methods.

Figure 6.2: (Left) Optical-flow features for detecting trains in motion. (Right) Representing people: ellipses (for tracking) and points (for articulation features).

The audio detector uses these two properties as input for a decision tree. An example of normal and aggressive speech can be seen in figure 6.1.

6.1.2 Video unit

Analysis of the video stream aims primarily at computing visual cues characteristic for physical aggression among humans. Physical aggression is usually characterized by fast articulation of body parts (i.e. arms, legs). Therefore, a principled approach for detecting aggression involves detailed body-pose estimation, possibly in 3D, followed by ballistic analysis of movements of body parts. Unfortunately, at present pose estimation remains a significant computational challenge. Various approaches (Gavrila, 1999) operate at limited rates and handle mostly a single person in a constrained setting (limited occlusions, pre-fitted body model). Simplified approaches rely on a coarser representation human body. An example is a system (Datta et al., 2002) that tracks a head of a person by analyzing body contour and correlates aggression with head's "jerk" (derivative of acceleration). In practice, high-order derivatives related to body contours are difficult to estimate robustly in cases where the background is not static and there is a possibility of occlusion.

Visual aggression features

Here we consider alternative cues based on an intuitive observation that aggressive behavior leads to highly energetic body articulation. We estimate (pseudo-) kinetic energy of body parts using a "bag-of-points" body model. The approach relies on simple image processing operations and yields features highly correlated with aggression. For detecting people our video subsystem employs adaptive background/foreground subtraction technique (Zivkovic, 2004). The assumption of static background scene holds fairly well in the center view area, where most of the people enter the scene. After detection, people are represented as ellipses and tracked with an extended

version of the mean-shift tracking algorithm (Zivkovic and Kröse, 2004). The extended tracker adapts position and shape of the ellipse (tilt, axes) and thus facilitates a close approximation of body area even for tilted/bended poses. Additionally, the mean-shift tracker handles well partial occlusions and achieves near real-time performance. We consider human body as a collection of loosely connected points with identical mass. While such a model is clearly a simplification, it reflects well the non-rigid nature of a body and facilitates fast computations. Assuming Q points attached to various body-parts, the average kinetic energy of an articulating body is given by the average kinetic energy of points,

$$E = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{2}m_i|v_i - v_e|^2 \tag{6.1}$$

where v_i, m_i denote, respectively, velocity vector and mass of the i-th point, and v_e denotes the velocity vector of the ellipse representing the person. By discounting overall body motion we capture only the articulation energy. Due to the assumption of uniform mass distribution between points, the total body mass becomes a scale factor. By omitting scale factors, we obtain a *pseudo*-kinetic energy estimate in the form $E = \frac{1}{Q}\sum_{i=1}^{Q}|v_i - v_e|^2$ Such features are assumed to measure articulation and will be our primary visual cues for aggression detection. Computation of energy features requires selecting image points that represent a person. Ideally, the points would cover the limbs and the head since these parts are mostly articulated. Further, to estimate velocities, the selected points must be easy to track. Accordingly, we select Q = 100 points within an extended bounding box of a person by finding pixels with the most local contrast (Shi and Tomasi, 1994). Such points are easy to track and usually align well with edges in an image (which in turn often coincide with limbs as in figure 6.2, right). For point tracking we use the KLT algorithm (Shi and Tomasi, 1994) (freely available implementation from the OpenCV library).

Train detection

An additional objective of the video unit is detecting trains in motion. Trains moving in and out of a station produce sonic noise that often leads to spurious audio-aggression detections. Therefore recognizing trains in video opens a possibility for later suppressing of such detections. A train usually appears as a large, rigid body and moves along a constrained trajectory. For a given view and rail section we define a mask that indicates the image regions where a train typically appears. In this region we track frame-to-frame motion of N = 100 image features with KLT (Shi and Tomasi, 1994) tracker (figure 6.2, left). The features' motion vectors are classified as train/non-train by a pre-trained nearest neighbor classifier. A train in motion is de-

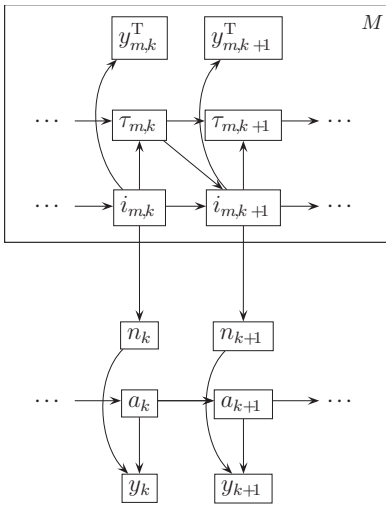


Figure 6.3: Dynamic Bayesian Network representing the probabilistic fusion model. The rectangular plate indicates M = 4 replications of the train sub-network.

tected when more then 50% of the features are classified positively. Due to the constrained movement of trains, our detector turns out quite robust to occasional occlusions of the train area by people.

6.1.3 Fusion unit

The fusion unit produces an aggregate aggression indication given the features/events produced independently by the audio and video subsystems. A fundamental difficulty with fusion arises from inevitable ambiguities in human behavior which make it difficult to separate normal from aggressive activities (even for a human observer). Additional problems follow from various noise artifacts in the sensory data. Given the noisy and ambiguous domain we resort to a probabilistic formulation. The fusion unit employs a probabilistic time-series model (a Dynamic Bayesian Network, DBN (Ghahramani, 2001)), where aggression level can be estimated in a principled way by solving appropriate inference problem.

Basic model

We denote the discrete-time index as $k = 1, 2, ...$, and set the gap between discrete-time steps (clock ticks) to 50ms. At the k-th step, $y^a_k \in 0, 1$ denotes the output of audio aggression detector, and $y^v_{j,k}$ denotes the pseudo-kinetic energy of the j-th, $j = 1, ..., J$, person. Our system can comprise several train detectors monitoring non-overlapping rail sections. The binary output of the m-th, $m = 1, ..., 4 = M$, train detector will be denoted as $y^T_{m,k} \in 0, 1$. (We tested a configuration with $M = 4$.)

Our aim is to detect "ambient" scene aggression, without deciding precisely which persons are aggressive. Therefore we reason on the basis of a cumulative articulation measurement $y_k^v = \sum_j y_{j,k}^v$ over all persons. Additionally, the cumulative is quite robust to (near-)occlusions when articulation of one person could be wrongly attributed to another.

In order to reason about aggression level, we use a 5-step discrete scale $< 0, 1 >$: 0.0 (no activity), 0.2 (normal activity), 0.4 (attention required), 0.6 (minor disturbance), 0.8 (major disturbance), and 1.0 (critical aggression).

Importantly, the aggression level obeys specific correlations over time and should be represented as a process (rather than an instantaneous quantity). We will denote aggression level at step k as a_k and define a stochastic process a_k with dynamics given by a 1st order model:

$$p(a_{k+1} = i | a_k = j) \quad \mathrm{CPT}_a(i, j) \tag{6.2}$$

where $\mathrm{CPT}_a(i, j)$, denotes a conditional probability table. In a sense, the first order model is a simplification as it captures only short-term dependencies.

The measured visual (y_k^v) and auditory (y_k^a) features are treated as samples from an observation distribution (model) that depends on the aggression level a_k. Since (later on) we will incorporate information about passing trains, we introduce a latent train-noise indicator variable $n_k \in 0, 1$ and assume that the observation model

$$p(y_k^v, y_k^a) | a_k, n_k) \tag{6.3}$$

depends also on the train-noise indicator. The model takes the form of a conditional probability table CPT_o, where the cumulative articulation feature is discretized.

Train models

The fusion DBN comprises several subnetworks — train models which couple train detections $y_{m,k}^T$ with the latent train-noise indicator n_k. Additionally, each train model encodes prior information about duration of a train pass. For the m-th rail section, we introduce a latent indicator $i_{m,k} \in 0, 1$ of a train passing at step k. We assume that the train detections $y_{m,k}^T$, the train-pass indicators $i_{m,k}$, and the train noise n_k obey a probabilistic relation

$$p(y_{m,k}^T | i_{m,k}) = \mathrm{CPT}_t(y_{m,k}^T, i^{m,k}) \tag{6.4}$$

$$p(n_k | i_{1:M,k}) = \mathrm{CPT}_n(n_k, i_{1:M,k}) \tag{6.5}$$

For each rail, the model 6.4 encodes inaccuracies of detector (mis-detections, false alarms). The model 6.5 represents the fact that passing trains usually induce noise, but also that sometimes noise is present without a passing train. Since a typical pass takes $5 - 10$ seconds ($100 - 200$ steps) the pass

indicator variable exhibits strong temporal correlations. We represent such correlations with a time-series model based on a gamma distribution. A gamma pdf $\gamma(\tau_m, \alpha_m, \beta_m)$ is a convenient choice for modeling duration τ_m of an event (α_m, β_m are parameters). To apply this model in a time-series formulation, we replace the total duration τ_m with a partial duration $\tau_{m,k}$ that indicates how long a train is already passing a scene at step k. By considering a joint process $i_{m,k}, \tau_{m,k}$ temporal correlations can be enforced by the following model

$$p \quad (i_{m,k+1} = 1 | \tau_{m,k}, i_{m,k} = 0) \quad = \eta_m \tag{6.6}$$

$$p \quad (i_{m,k+1} = 1 | \tau_{m,k}, i_{m,k} = 1) \quad = p(\tau_m > \tau_{m,k}) = \tag{6.7}$$

$$= \int_{\tau_{m,k}}^{+\infty} \gamma(\tau_m, \alpha_m, \beta_m) d\tau_m \quad = 1 - F(\tau_{m,k}, \alpha_m, \beta_m) \tag{6.8}$$

where $F()$ is a gamma cumulative density function. Parameter η_m denotes a probability of starting a new train pass. At the k-th step, the probability of continuing a pass is function of the current duration of the pass. A configuration ($i_{m,k+1} = 1, \tau_{m,k}, i_{m,k} = 1$) implies that a pass does not finish yet and the total pass duration will be larger than $\tau_{m,k}$, hence the integration. Further, the partial duration variable obeys a deterministic regime

$$\tau_{m,k+1} = \begin{cases} 0 & \text{iff} \quad i_{m,k+1} = 0 \\ \tau_{m,k+1} = \tau m, k + \epsilon & \text{otherwise} \end{cases} \tag{6.9}$$

where ϵ = 50 ms is the period between successive steps.

Inference and learning

In a probabilistic framework, reasoning about aggression corresponds to solving probabilistic inference problems. In an online mode, the key quantity of interest is the posterior distribution on aggression level given data collected at up to the current step, $p(a_k | y^v_{1:k}, y^a_{1:k}, y^t_{1:m,1:k})$. From this distribution we calculate the expected aggression value, which will be the basic output of the fusion system.

Given the graphical structure of the model (figure 6.3), the required distribution can be efficiently computed using a recursive, forward filtering procedure (Ghahramani, 2001). We implemented an approximate variant of the filtering procedure, known as the Boyen-Koller algorithm (Boyen and Koller, 1998). At a given step k, the algorithm maintains only marginal distributions $p(h_k | y^v_{1:k}, y^a_{1:k}, y^t_{1:m,1:k})$, where h_k is any of the latent variables. When new detector data arrive, the current-step marginals are updated to represent the next-step marginals. An important modeling aspect are temporal developments of processes in the scene. Unlike the binary train-pass events, aggression level usually undergoes more subtle evolutions as the tension and anger among people build up. Since the assumed (1st-order) model might

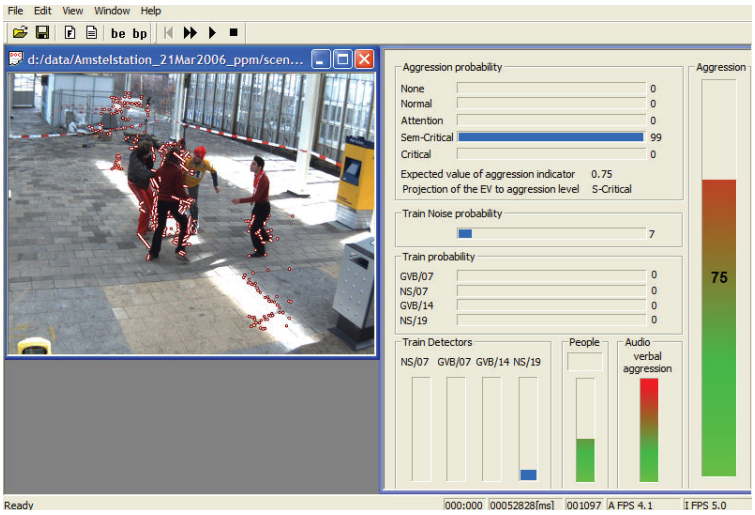


Figure 6.4: A screen-shot of the CASSANDRA prototype system. In the lower part of the right window, various intermediate quantities are shown: the probabilities for trains on various tracks (currently zero), the output of the video-based articulation energy ("People", currently mid-way), the output of the audio detector (currently at maximum). The slider covering the right side shows the overall estimated aggression level (currently "major disturbance").

not capture well long-term effects and a stronger model would be rather complicated we enforce temporal correlations with a simple low-pass filter. The articulation measurements (before inference) and the expected aggression level (after inference) are low-pass filtered using a 10 s running-average filter. The parameters of our model: probability tables: CPT_a, CPT_o, CPT_n, CPT_t and the parameters α_m, β_m of the gamma pdf's) are estimated by maximum-likelihood learning. The learning process relies on detector measurements from training audio-video clips and ground-truth annotations. The annotations are particularly important for learning the observation model CPT_o. An increased probability of a false auditory aggression in the presence of train noise, will suppress the contribution of audio data to the aggression level when the video subsystem reports a passing train.

6.2 Experiments

We evaluate the aggression detection system using a set of 13 audio-video clips (scenarios) recorded at a train station. The clips (each 100s-150s) feature 2-4 professional actors who engage in a variety of activities ranging from normal (walking) through slightly excited (shouting, running, hugging), moderate aggressive (pushing, hitting a vending machine) to critically aggressive (football-supporters clashing). The recording took place at a platform of an

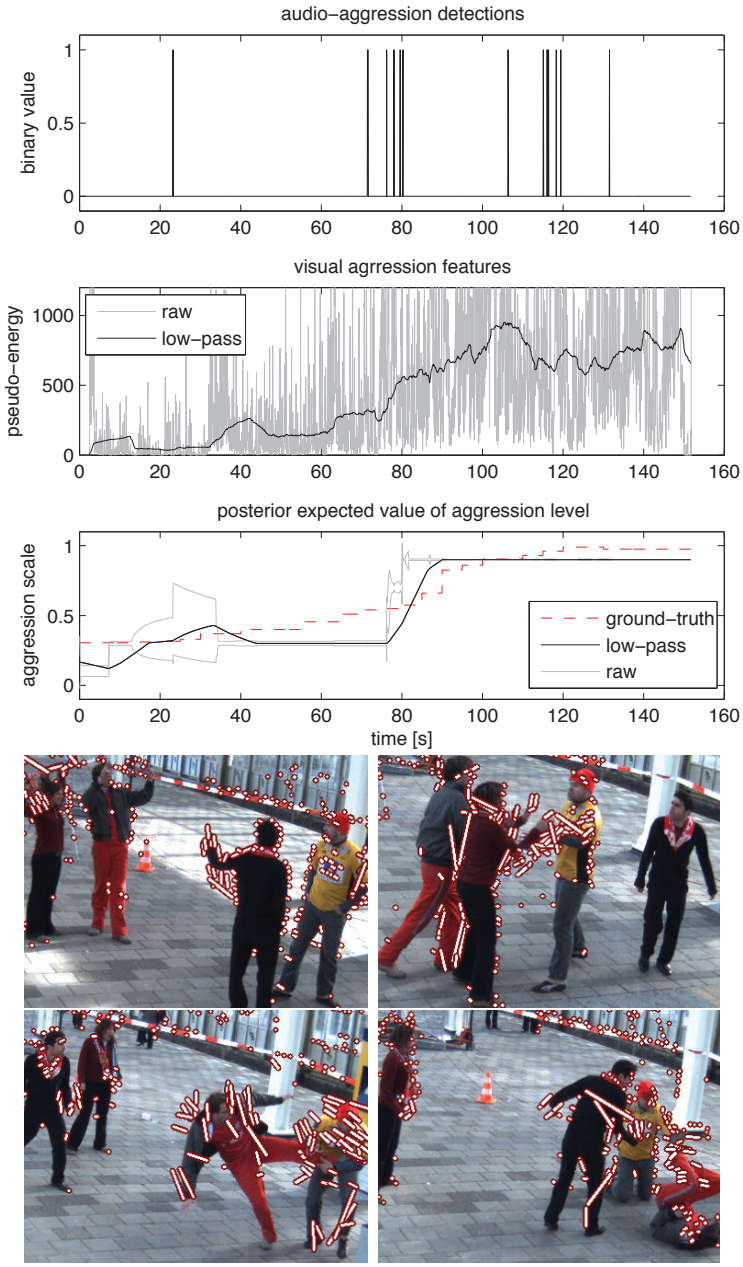


Figure 6.5: Aggression build-up scenario. (Top graph) Audio-aggression detections. (Middle graph) Visual articulation measurements. (Bottom graph) Estimated and ground-truth aggression level. The gray lines show uncertainty intervals (2ÃŮ std. deviation) around raw (before filtering) expected level. (Images) Several frames (timestamps: 45 s, 77 s, 91 s, 95 s). Notice correspondence with the articulation measurements.

actual train station (between two rail tracks, partially outdoor) and therefore incorporates realistic artifacts, like noise and vibrations from trains, variable illumination, wind, etc. Scenarios have been manually annotated with a ground-truth aggression level by two independent observers using the scale mentioned in section 6.1.3. Aggression toward objects was rated approx. 25% lower than aggression toward humans, i.e. the former did not exceed a level of 0.8. Figure 6.5 details the results of the CASSANDRA system on a scenario involving gradual aggression build-up. Here, two pairs of competing supporters first start arguing, then get in a fight. The bottom panel shows some illustrative frames, with articulation features highlighted. We see from figure 6.5 that the raw (before low-pass filtering) articulation measurements are rather noisy, however the low-pass filtering reveals strong correlation with ground-truth aggression level. The effect of low-pass filtering of the estimated aggression level is shown bottom plot of figure 6.5. A screen-shot of the CASSANDRA system in action is shown in figure 6.4. We considered two quantitative criteria to evaluate our system. The first is the deviation of the CASSANDRA estimated aggression level from the ground-truth annotation. Here we obtained a deviation of mean 0.17 with standard deviation 0.1. The second performance criterion considers aggression detection as a two-class classification problem of distinguishing between "normal" and "aggressive" events (by thresholding aggression level at 0.5). Matching ground-truth with estimated events allows us to compute detection rate (%) and false-alarm rate (per hour). When matching events we allowed a time deviation of 10 s. The cumulative results of leave-one-out tests on 13 scenarios (12 for training, 1 for testing) are given in figure 6.7. Comparing the test results for three modalities (audio, video, fusion of audio+ video), we notice that the auditory and visual features indeed are complimentary; with fusion the overall detection rate increased without introducing additional false alarms. It is important to note that our dataset is heavily biased toward occurrences of aggression, i.e. which put the system to a difficult test. We expect CASSANDRA to produce much less false alarms in a typical surveillance setting, where most of the time nothing happens. Table 6.1 gives an overview of the detection results on the scenarios. We notice that the system performed well on the clearly normal cases (scenarios 1-3) or aggressive cases (scenarios 9-13), while borderline scenarios were more difficult to classify. The borderline behavior (e.g. scenarios 7-8) turns out also difficult to classify for human observers given the inconsistent ground-truth annotation in Tab. 1. The CASSANDRA system runs on two PCs (one with the video and fusion units, the other with the audio unit). The overall processing rate is approx. 5Hz with 756x560x 20Hz input video stream and 44 kHz input audio stream.

Table 6.1: Aggression detection results by scenario. The table indicates number of events (positive = aggressive). Figure 6.6 shows example frames from the 5th scenario.

		ground-truth	detected events	
id	scenario content	positive	true-pos.	false-pos.
1	normal: walking, greeting	0	0	0
2	normal: walking, greeting	0	0	1
3	excited: lively argument	0	0	0
4	excited: lively argument	1	1	0
5	aggression toward a vend. machine	1	0	1
6	aggression toward a vend. machine	1	0	0
7	happy football supporters	1	1	0
8	happy football supporters	0	0	1
9	supporters harassing a passenger	1	1	0
10	supporters harassing a passenger	1	1	0
11	two people fight, third intervenes	1	1	0
12	four people fighting	1	1	0
13	four people fighting	1	1	1

Figure 6.6: Selected frames from a scenario involving aggression toward a machine.

6.3 Conclusions and future work

We demonstrated a prototype system that uses a Dynamical Bayesian Network to fuse auditory and visual information for detecting aggressive behavior. On the auditory side, the system relies on scream-like cues, and on the video side, the system uses motion features related to articulation. We obtained a promising aggression detection performance in a complex, real-world train station setting, operating in near realtime. The present system is able to distinguish well between clear cases of aggression and normal behavior. In the future, we plan to focus increasingly on the "borderline" cases. For this, we expect to use more elaborate auditory cues (laughter vs scream), more detailed visual cues (indications of body-contact, partial body-pose estimation), and stronger use of context information.

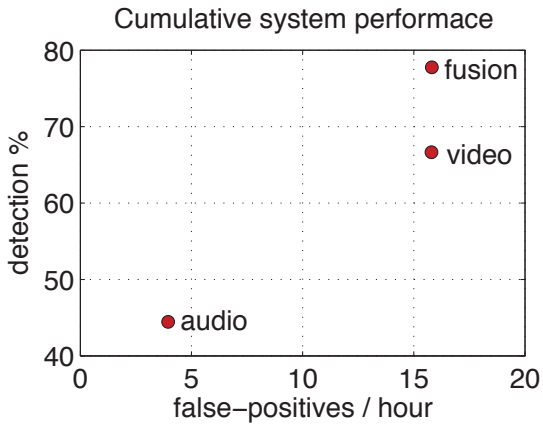


Figure 6.7: Cumulative detection results by sensor modality.

7

Evaluation on a dataset of city sounds

This chapter is based on: Maria E. Niessen, Johannes D Krijnders, and Tjeerd C Andringa. Understanding a soundscape through its components. In *Proceedings of Euronoise 2009*, EN09_242, 2009.

In chapter 4 we developed a recognition technique based on signal-driven recognition and knowledge-driven re-evaluation. This technique was tested on a dataset recorded on a station in the Netherlands, while this was an uncontrolled environment, recordings were only made during one day and the number of possible sound classes was limited. In this chapter we present the application of the same techniques on a more diverse dataset. Recordings were made on several different days and with changing weather conditions. In addition the recordings were made on six different locations. The result was a dataset that is more representative for a system employed in a city.

In the following section we will describe the methods that we developed to segment and label components in a soundscape. In the third section we present a dataset, which is used in an experiment to test the combined methods. Finally we will discuss the results of the experiment, and give an outlook on future work.

7.1 Methods

To identify acoustic events in a continuous sound signal, we first selected components from the sound signal that are likely to stem from a single source (section 7.1.1). Subsequently, we applied a model of human memory to select the most likely label for the events that constitute these components, based on a prediction of the recording location (section 7.1.2). The methods are only briefly described here. For a detailed description we refer to chapter 3 (Krijnders et al., 2010) and chapter 4 (Niessen et al., 2009b).

7.1.1 Sound Processing

The spectrogram of the sound signal is segmented on the basis of the local spectro-temporal properties. Segments are likely to stem from a single source when they are based on local properties. For example, local energy maxima that resemble tones and are developing smoothly in time are likely to stem from the same source. The robustness and reliability of these segments, called signal components, are improved with grouping principles from auditory scene analysis, such as common onset, common offset and common frequency development (Bregman, 1990; Ellis, 1999). The strategy to combine local signal properties and grouping principles allows the selection of qualitatively different types of groups, namely tones and harmonic complexes, pulses, and broadband events. A physical description of these groups is used to classify and label them as sound events using a k-nearest neighbor (k-NN) classifier.

7.1.2 Dynamic Network Model

The segmented and labeled group derived in the previous section are reevaluated using a dynamic network model. This is beneficial because the sound signal can be distorted or masked by transmission effects. Furthermore, the labeling using a k-NN classifier is never perfect. To solve these problem the network incorporates knowledge of the context, based on the short-term history of events.

The network is identical to the one used in section 4.2 (Niessen, 2010), with the exception that the top-level nodes represent recording locations rather than sequences of events. This is more appropriate for this dataset and it allows the network to predict the location of the recording.

7.2 Experiment

We present an experiment to demonstrate that the proposed methods can be used to identify events in a soundscape given a predicted location. First, we

Table 7.1: Examples of annotated classes and their occurrences.

Class	Total number of occurrences	Sum of duration of occurrences
Bird	238	17 min
Bike	30	2 min 20 sec
Rooster	16	43 sec
Horn	8	11 sec
Shopping bag	1	7 sec

describe the dataset that is used in the experiment. Next, the setup of the experiment is explained, and in the last part we present the results of the experiment.

7.2.1 Data

The dataset was collected under different weather conditions on a number of days in March 2009 in the town of Assen (65,000 inhabitants, in the north of the Netherlands). The recordings were made by six groups of three students as part of a master course on sound recognition. Each group made recordings of three minutes at six different locations: a railway station platform, a pedestrian crossing with traffic lights, a small park-like square, a pedestrian shopping area, the edge of a forest near a cemetery, and a walk between two of the positions. Recordings were made using M-Audio Microtrack-II recorders with the supplied stereo microphone at 48 kHz and 24 bits stereo. This data, with annotations, will be made available on http://daresounds.org.

All the recordings were annotated by two students separately. These two annotations were merged, such that equal labels did not overlap, but became one instance. We examined the resulting merged annotations, and adjusted them when necessary. However, we did not introduce new annotations. (An exception was made for the annotations of one group, which we had to complete because they were too meager.) We ensured that the names of events were uniform across all the files to prevent the dynamic network model from learning annotators rather than locations. The total of 44 audio files, with an average duration of 3,5 minutes, were annotated for 54 different classes. However, half of these classes were annotated less than 5 times, while just a few classes comprised most of the annotations. In table 7.1 a few examples of annotated classes are given, ranked according to their frequency in the complete dataset.

7.2.2 Setup

The annotations of sound recordings were used to train both the k-NN classifier, for the labeling of the signal-driven groups, and the knowledge of the dynamic network. For the k-NN classifier, all 44 audio files were processed

with the signal-driven method described in section 7.1.1. The segmented groups with the highest score that overlapped with an annotation were given that annotation as a label. Groups that did not overlap in time with an annotation were labeled as noise. All other groups were discarded. From these processed files, 44 file pairs were generated. Each file pair consisted of a file used for training, for which the labeled groups from 43 files were used, and a test file, which contained all the groups from the one file that was left out, resulting in a leave-one-out set. Additionally, the annotations of the training file of each file pair were used to train the weights in the dynamic network model (see section 7.1.2).

In the test phase, the groups in the test file are used as input for the dynamic network (see figure 4.3). Subsequently, the possible classes that the group can represent are initiated as event hypotheses in the network. The weight between the group and the event hypotheses is the probability of each class given by the k-NN classifier. If the event cannot be classified and is labeled as noise by the k-NN classifier, the weights are set to the prior probabilities of each event. Based on the events hypothesized by the network, the network forms a hypothesis of the location, which in turn initiates expectancies of certain sound events that might follow. The results of this combined approach are the most likely events that explain the segmented groups, given the identified sound events and their predicted location.

The most likely events, according to the k-NN classifier and the combined model, are compared to the annotations to measure the performance of the recognition system. The performance is measured with the F-measure. The F-measure is used in information retrieval to test the effectiveness of the performance of a system (van Rijsbergen, 1979), for example a search engine. The F-measure is computed as the harmonic mean between the recall, which represents whether relevant results are retrieved, and the precision, which represents whether irrelevant results are not retrieved. Applied to the results of automatic sound identification, precision is a measure for the fraction of time the identifications are correct, and recall is a measure for the fraction of identifications that are made out of the amount that should have made.

7.2.3 Results

The success of the dynamic network model, as applied in this study, depends on whether the location prediction is correct. The location predictions of the test files are listed in table 7.2. The number of test files at each location is included between parentheses behind the location name. The location predictions of the 7 test files of recordings while walking are not included, because they cannot be assigned to a single location. The top 1 indicates how many location predictions are correct on average for a specific location (the spread in standard deviations is given between parentheses). The model has

Table 7.2: Correct classification rate (average and spread) of location predictions.

Location	Top 1	Top 2	Top 3
City center (7)	0.01 (0.02)	0.02 (0.02)	0.03 (0.03)
Graveyard (7)	0.04 (0.05)	0.04 (0.05)	0.20 (0.16)
Museum (8)	0.25 (0.16)	0.78 (0.30)	0.89 (0.16)
Traffic lights (7)	0.04 (0.05)	0.18 (0.19)	0.65 (0.29)
Train station (8)	0.66 (0.19)	0.85 (0.07)	0.88 (0.07)

an activation or confidence value for all the location hypotheses. Therefore, if the best prediction is not correct, the second best might be. The top 2 and 3 specify whether the correct location is among the second or third best predictions.

Only two locations can be predicted well, the train station and the museum, because some of the sounds the model can identify are very specific for one of these two locations, such as train sounds for the train station. In contrast, many of the other sounds the model can identify reliably, such as cars and speech, are generic, and can occur at any of the locations. Therefore, the location prediction is not reliable in many test files and also not crucial as predictions would the same anyway.

The location prediction is based on the classified segmented groups, and used to select the most likely label for the group. Of all 54 annotated classes, 12 classes are identified (segmented and labeled) by the combined model (the segmentation algorithm, the k-NN classifier, and the dynamic network). These 12 classes are the classes that are often annotated. Hence, the k-NN classifier and the dynamic network model can learn these classes better than classes that occur infrequently. Table 3 shows the F-measure, precision and recall of the identifications made by the k-NN classifier (K) and by the dynamic network model (D) for the 12 classes. The number of test files (out of the total of 44) in which at least one instance of a class was found by either one of the models, is given in parentheses behind the class name. The F-measures that are zero for both models are not included in the mean values in the table. The bottom row indicates the measures weighted for the number of test files. On average, the dynamic network model improves the F-measure, mostly through an increased recall, which means that more correct instances of annotations are recognized than with the k-NN classifier in isolation.

7.3 Conclusions

In the previous section we have demonstrated that a model that combines both signal-driven algorithms and knowledge in the form of the predicted location, improves the identification of sound events in a real-world environ-

Table 7.3: F-measure, precision, and recall of k-NN classifier (K) and dynamic network model (D).

Sound class	F-measure (K / D)	Precision (K / D)	Recall (K / D)
Bird (6)	0.02 / 0.17	0.34 / 0.65	0.01 / 0.12
Braking train (3)	0.20 / 0.09	0.15 / 0.12	0.32 / 0.14
Bus (8)	0.10 / 0.23	0.22 / 0.19	0.09 / 0.41
Car (27)	0.45 / 0.36	0.62 / 0.53	0.43 / 0.35
Footsteps (12)	0.02 / 0.17	0.49 / 0.71	0.01 / 0.14
Passing train (2)	0.73 / 0.73	0.62 / 0.62	1 / 1
Pressure cleaning (1)	0 / 0.08	0 / 1	0 / 0.04
Speech (15)	0 / 0.08	0.52 / 0.33	0.02 / 0.16
Starting train (2)	0.22 / 0.13	1 / 0.50	0.12 / 0.08
Truck (3)	0.05 / 0.25	0.63 / 0.88	0.02 / 0.15
Truck stationary (1)	0.09 / 0	1 / 0	0.05 / 0
Wind (24)	0.10 / 0.19	0.48 / 0.34	0.08 / 0.25
Weighted average	0.18 / 0.24 (+33%)	0.50 / 0.46 (-8%)	0.17 / 0.26 (+53%)

ment. The overall results may not seem impressive, but this is partly explained by the performance measure. The F-measure is based on the overlap of the annotations and the labeled groups. Therefore, it is dependent on both the annotations and the detection algorithm. Annotating sound is a complex process. The annotators did not only use information in the recording, but also knowledge of the environment, because they were present during the recordings. We cannot determine to what extend the annotations are based on the recording or on their knowledge. Some annotated sound events can even hardly be identified by a human listener who has to rely on the audio signal alone.

In contrast, the detection algorithm relies only on the recorded signal. This signal is uncontrolled and thus very challenging for the algorithm that segments relevant parts. Furthermore, the recordings contain a wide variety of sounds events, most of which occur only a few times in all the recordings. To be able to learn the patterns of a sound event, the k-NN classifier (or any other classifier) needs more examples than were available of most classes in the dataset in this study.

These observations indicate that modeling the context is essential to achieve robust event identification in real-world environments. Indeed we have shown that context, in the form of location, improves event identification substantially, even though it is so far only based on acoustic information. Since the dynamic network model relies on the segmented groups, it cannot identify events that are not segmented. Additionally, the location is not predictive for many generic classes, such as speech and cars. However, the generic classes occur most often, and are best classified by the k-NN classifier. In other words, the rare events are the events that are good predictors of a location, while these are the hardest events to learn, because

they are rare. However, the dynamic network model is not limited to process acoustic information. In another study we show that the dynamic network model can also be used to improve visual robot localization (Niessen et al., 2009a). Because the model can receive input from different modalities, it can combine multiple modalities and factors in a single system that returns a single analysis. We plan to integrate information from multiple sources of knowledge so that the context is modeled more profoundly.

In summary, the combined model provides a new way to analyze soundscapes by identifying its components. Because these components are also based on knowledge of the context of the acoustic events, they are a first approximation of meaningful events. However, to improve the identification of these components in the complexity of real soundscapes, we require a combined development of segmentation algorithms and models that can include non-acoustical factors. Furthermore, we will study human perception in parallel, so we can validate the model for soundscape analysis. Vice versa, the development of a system to analyze a soundscape automatically might increase our understanding of human soundscape perception.

8

Automatic Extraction of Formants in Noise

This chapter is based on: Bea Valkenier, Johannes D. Krijnders, R.A.J. van Elburg, and T.C. Andringa. Robust vowel detection. In *Proceedings of NAG/DAGA 2009*, pages 1306-1309, 2009. and Bea Valkenier, Johannes D. Krijnders, R.A.J. van Elburg, and T.C. Andringa. Automatic Extraction of Formants in Noise. Submitted to *Interspeech*, 2010.

8.1 Introduction

In previous chapters we have shown the application of the recognition techniques developed in chapters 3 and 4 on realistic recordings. While the application to realistic recordings is the main goal of these techniques it is useful to test the techniques on datasets that are carefully recorded and selected because the annotation are less ambiguous. In this chapter the dataset consists of vowels and is annotated both on formants and fundamental frequency. Formants are the resonance frequencies of the vocal tract; they change as the shape of the vocal tract changes. As such, formants are important acoustical cues for the description and identification of phonemes. The task of automatic formant frequency estimation is traditionally investigated by methods based on spectral analysis. Such representations can be used to accurately estimate the formant positions and formant develop-

ments Vargas and McLaughlin (2008) in clean speech. However, efforts that focus on formant detection in noise (Mustafa and Bruce, 2006; Hillenbrand and Houde, 2003; de Wet et al., 2004; Yan et al., 2006) show much worse performance.

One of the fundamental problems with spectral analysis is that signal and noise are treated alike and spectral shape information is spread over all parameters. As a result, the features are not stable through varying noise conditions. This urges the user to train and test a system in similar conditions. Furthermore the possibility to suppress noise or separate sources after feature estimation is reduced. Improved preprocessing yielded only limited progress towards a solution of this problem. For instance, cepstral mean subtraction can lead to acceptable recognition results (Yan et al., 2006) but only as long as the acoustic environment complies to highly specific conditions, such as predictable or steady noise. Other methods try to identify unreliable regions before recognition and analyze only the parts that are marked as reliable in order to bias the information towards representing the target speech Cooke et al. (2001). Such methods work fine as long as enough reliable observations are made which is not the case in SNR's lower than 0dB. Although these and other methods improve the signal descriptions in noisy conditions, the fundamental problem of inclusion of the noise in the spectral features is still unaddressed.

In contrast, human listeners can detect and recognize speech with relatively little hindrance of background noises (Lippmann, 1997; O'Shaughnessy, 2008) which might partly be explained by the characteristics of the extracted features. Human listeners might exploit the fact that some of the informative constituents of the speech sound, namely harmonics near formant positions are relatively robust to noise. Formants or equivalently resonances in the vocal tract stand out energetically and are robust to noise. As a result, the same or similar values could, in principle, be automatically derived from noisy as well as clean speech. In this paper, we test this assumption. We present a newly developed formant-detection algorithm, which can be implemented as a real-time system, that uses features similar to the features hypothesized to be used by humans. We test the robustness of the algorithm to noise and compare with the results from (Yan et al., 2006). Next we apply a simple classification method (Best First Search) in order to compare our results with results from (de Wet et al., 2004) who used the same database to test the robustness of formant-like features.

8.2 Method

8.2.1 Algorithm

To calculate the formant trajectories we calculate a cochleogram of the audio signal with the gamma-tone filterbank as described in section 2.3. For the cochleogram the tone-fit (chapter 3) is calculated and the from the tone-fit matrix signal-components are extracted (chapter 4). An example of signal components extracted from a cochleogram of the utterance "hud" can be seen in figure 8.1(a). These signal components are combined to form harmonic complexes using the methods described in section 4.1.3. The hypothesized harmonic complex with the highest score is used in the next step. The fundamental frequency and selected signal components can be seen in figure 8.1(b). Because not all harmonics are found as signal components the next stages of processing use the energy at the harmonic positions based on the fundamental frequency.

Except for a special case, Lombard speech (section 6.1.1, Junqua (1993)), the formant trajectories do not coincide with the harmonics. Therefore a quadratic interpolation is applied to estimate the real formant location from the harmonics around and including the maximum. This interpolation provides the final estimate of the formant positions as shown in figure 8.1(c). Formant estimates with minimal distance in the frequency plane are connected into formant tracks. Finally we keep formants of sufficient duration (7 frames or more, figure 8.1(d)).

8.2.2 Material

The formant extractor was tested on the American English Vowels dataset (AEV) (Hillenbrand et al., 1994). The dataset consists of 12 vowels pronounced in /h-V-d/ context by 48 female, 45 male and 46 child speakers. All vowels can be correctly classified by American English listeners. The AEV dataset is annotated for the first four formants at 8 points in time for each vowel, which makes it a suitable ground truth. We added pink noise in decreasing signal to noise ratios (SNRs), from 30dB to -14dB SNR. The step size was 10dB at SNR's above zero and 2dB at lower SNR's . Pink noise was chosen because it masks speech evenly.

8.2.3 Evaluation

Two performance measures on formant detection for the first three annotated formants are calculated. The first three formants are necessary to classify vowels . First, a detection ratio (r_d) is calculated, giving the fraction of

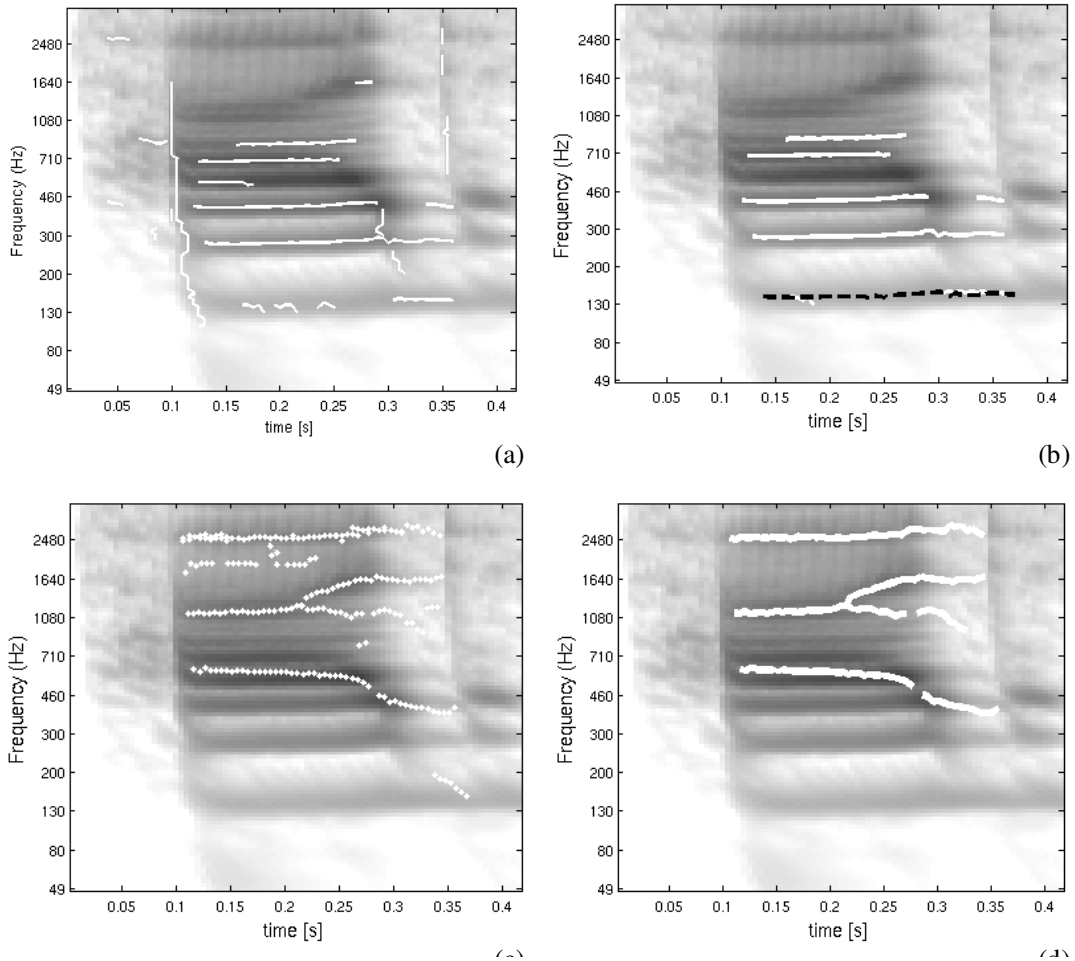


Figure 8.1: Cochleogram of a male speaker pronouncing [hud]. (a) Energetic signal components (b) selected HC, the fundamental frequency is given by the striped line (c) formant detections (d) selected formants.

annotated formants that is consistent with our detections,

$$r_d = \frac{\#(detected \bigcap annotated)}{\#(annotated)} \tag{8.1}$$

We consider a detection to be consistent with the annotation if it falls within the range of 15% (1st formant), 12% (2nd formant) and 8% (3rd formant) from the annotated formant frequency. This equals a mean accepted error of respectively 95Hz, 316Hz and 266Hz. The range is chosen such that formants that were considered correct by the authors according to visual inspection were included. Second, a measure is calculated for the detected formants that cannot be related to the annotated formants, the spurious peaks (r_{sp}). This measure is the ratio between the number of extra detected

formants at the annotated positions, and the number of annotated points,

$$r_{sp} = \frac{\#(detected) - \#(detected \bigcap annotated)}{\#(annotated)} \tag{8.2}$$

Subsequently, the detected formants that are analogous to the ground truth formants are further investigated in terms of how well they are able to classify the vowels in the test material. To that end, a feature vector is constructed, consisting of the frequency values of only the subset of detected formants that are analogous to the reference formants. Due to missing values, i.e. formants that were not detected, we were limited to a small number of classification algorithms to choose from. The best first tree (BFT) search algorithm from the WEKA toolbox Witten and Frank (2005) allows a weighting of different features. This is a relevant characteristic because different formants represent a different informational value and should be weighted accordingly. We used the BFT search algorithm using a tenfold cross validation method on the detected formants.

8.3 Results

Two performance measures on formant detection for the first three annotated formants are calculated. The first three formants are necessary to classify vowels. First, a detection ratio (r_d) is calculated, giving the fraction of annotated formants that is consistent with our detections,

8.3.1 Formant extraction

In figure 8.2 (top left and top right, bottom left) the detection rates (r_d) and proportion of spurious peaks (r_{sp}) are plotted against an increasing SNR. In clean conditions, 90% correct detections are made for all three speaker classes for the first formant, and 75% correct for the second and third formants. For a 0dB SNR, this is reduced for female and child speakers to 70% for the first formant and 53% for the second formant; for male speakers the decline is steeper: for a 0 dB SNR, an r_d of 50% for the first and 35% for the second formant is found. For all three speaker classes formants consistent with the ground truth can still be extracted at negative SNR values. The proportion of spurious peaks increases gradually to 15% at 0dB for male speakers and -3dB for female and child speakers. To improve the interpretation of the results of the HC extraction stage, table 8.2 shows the occurrences of HC's that are not detected and the occurrences of HC's that exhibit an octave error compared to the fundamental frequency annotations in Hillenbrand et al. (1994). The percentage of not extracted, as well as wrongly extracted HC's, is much higher for male speakers than for female and child speakers which explains the relatively low performance of our method for

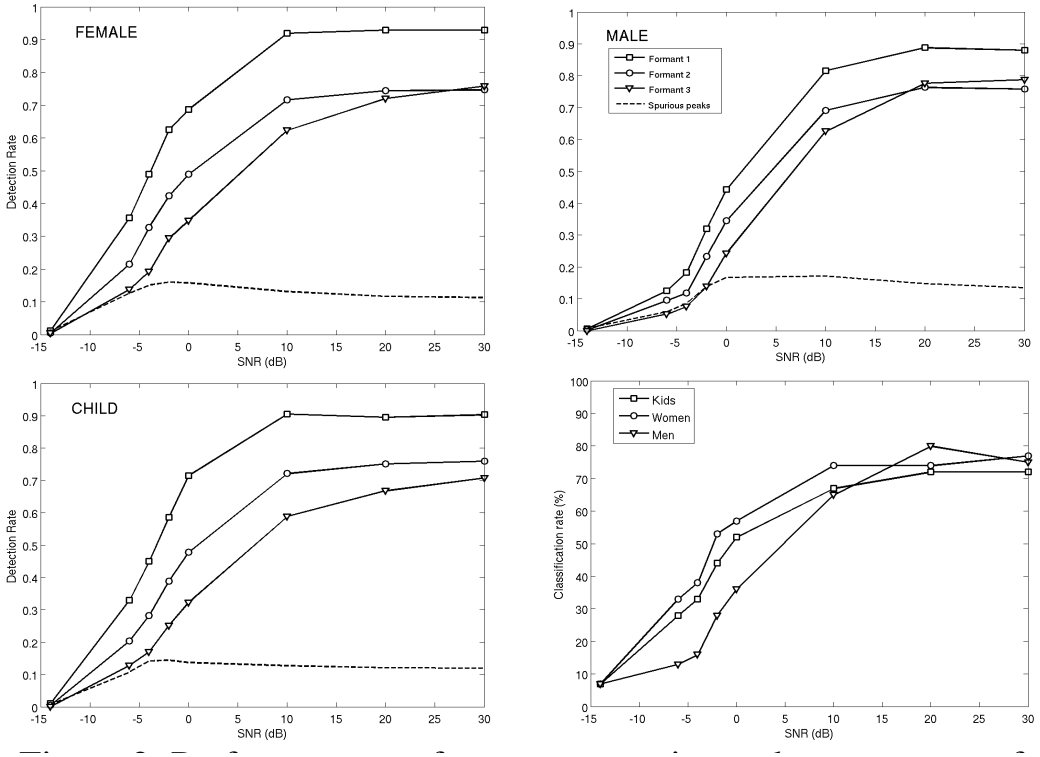


Figure 8.2: Performance on formant extraction task; percentage of correctly extracted formants according to the annotations for female (top left), male (top right) and child (bottom left) speakers. The bottom right panel gives classification results using a BFT search algorithm.

male speakers. The fact that even in clean conditions some HC's are missed or suffered from octave errors indicates that the criterion on harmonic relations of the tonal components is too strict. Therefore we expect a possible improvement by relaxing this criterion. In the r_{sp} measure a peak exists at low SNRs for all three speaker classes (figure 8.2, dotted line). This effect is due to an increased number of octave errors in noisy conditions. A fundamental frequency that is too low results in more harmonics between two formant positions which explains the relatively high amount of spurious peaks. If the SNR decreases, the number of incorrectly extracted harmonic complexes increases, which results in an increased amount of incorrect formant detections. If the SNR decreases further, the overall number of extracted harmonic complexes decreases due to missed harmonic complexes.

8.3.2 Vowel classification

The bottom right panel of figure 8.2 shows the classification scores obtained with the BFT search algorithm. Recognition in clean speech is 80% for

Table 8.1: Confusion matrix of classification task in clean speech pooled over all speaker classes. The background indicates the confusion found by Hillenbrand in a recognition experiments by human listeners, black represents 100%, white 0%

	ae	ah	aw	eh	ei	er	ih	iy	oa	oo	uh	uw
ae	92	3	2	35	1	0	2	0	0	0	4	0
ah	34	66	16	11	0	1	1	0	0	1	8	1
aw	0	21	96	0	0	1	1	0	0	5	13	2
eh	37	2	3	86	0	0	3	0	0	5	3	0
ei	1	0	0	0	115	1	12	9	0	0	0	1
er	0	0	2	0	3	129	1	0	1	3	0	0
ih	2	0	0	4	9	3	114	3	0	2	0	2
iy	0	1	0	0	11	2	5	114	1	3	0	2
oa	0	0	3	0	0	1	0	1	106	14	3	11
oo	0	2	0	1	0	5	0	0	17	96	4	14
uh	0	10	19	0	1	1	1	0	4	13	89	1
uw	0	0	0	0	0	2	3	1	4	14	0	115

all three speaker classes. In 0dB SNR, recognition for female speakers is 60% and recognition for male speakers 35%. Table 8.1 shows the confusion matrix of the classifications of the vowels from all speakers pooled together. Relatively many confusions occur between the vowel sounds ‘ae’, ‘eh’ and ‘ah’, ‘aw’. Those four vowels are confused with one of the other sounds for 25% percent of the vowels. It is noteworthy that the same vowels are reported to be confused most often by human listeners (Hillenbrand et al., 1994), see table 8.1.

8.4 Discussion

We described and tested a method to automatically extract formants based on robust parts of the acoustical signal, namely the harmonics at formant positions. In contrast to commonly used ASR features that degrade slowly as a function of decreasing SNR, formant positions remain constant under different noise conditions. The robustness of harmonics at formant positions allows us to develop a method to extract similar feature values over a range of acoustical conditions.

8.4.1 Harmonic complex extraction

The results in table 8.2 show that in clean situation the harmonic complex extraction works near perfect for female and child voices. This performance only starts to drop around 0 dB. The scoring function (equation 4.1) was not optimized for this dataset, but rather for the dataset used in chapter 4. This leaves room for improvement. The performance for male speakers, with a lower fundamental frequency, leaves room for future work as well.

Table 8.2: Type of mismatch for detection of the harmonic complex for male, female and child speakers. For male speakers more harmonic complexes are missed and more octave errors are made.

SNR (dB)	30	20	10	0	-2	-4	-6	-14
Female Not extracted(%)	0	0	1	18	23	35	51	98
Octave error(%)	1	2	3	10	11	13	11	1
Male Not extracted(%)	2	1	8	41	57	74	81	100
Octave error(%)	8	6	10	7	5	3	3	0
Child Not extracted(%)	0	2	1	17	28	39	51	98
Octave error(%)	1	1	2	4	6	9	8	1

8.4.2 Formant extraction

In noisy conditions our method compares well with existing methods proposed in the literature. For instance the method proposed by (Yan et al., 2006) is a linear prediction model consisting of a noise reduction stage, a secondary hidden Markov model (HMM2) and a Kalman filter. This method results in average estimation errors of respectively 17%, 12% and 8% for the first, second and third formants in a SNR of 0dB train noise. Although the method in (Yan et al., 2006) outperforms our method, a problem of the method proposed by (Yan et al., 2006) is that it cannot be easily generalized to unseen types of noise, as the noise reduction methods are specifically suitable for relatively stable types of noise and the method relies predominantly on de-noising of the input signal. In contrast to this our method can be generalized to all types of noise.

8.4.3 Vowel classification

In noisy conditions the results for female and child speakers compare favorably with results found in the literature. (de Wet et al., 2004) report on a vowel classification task on the same AEV database, in which they used HMM2 to evaluate probabilities of both frequency and time. Using this method 55% correct classifications in 0dB babble noise are found for female and male speakers. For female speakers our method results in higher scores (60%) although we used a simple classification mechanism (BFT search). We already mentioned the possibility of improving the part of the algorithm that extracts the harmonic complex and it is possible that identification can be improved with a more advanced learning algorithm. This yields the possibility to obtain better results in noisy conditions for all speaker classes compared to those reported by (de Wet et al., 2004). Apart from this positive comparison, it is noteworthy that by using features similar to the features hypothesized to be used by humans, we find confusions similar to those of human listeners.

8.5 Conclusion

We showed that it is possible to develop an automatic method to extract formant feature values over a range of acoustical conditions that uses the robustness of harmonics at formant positions. For pink noise we showed that formants consistent with the ground truth can be extracted at low and even negative SNR-values. We expect performance enhancement by further optimizing our harmonic complex detection algorithm. These initial results seem to suggest that formants, thought to be important for humans in speech processing, can also constitute robust features for automatic vowel detection, and possible automatic speech recognition, systems.

9

Conclusions

Recent developments in soundscape research and systems for ambient awareness have shown a need for a new range of sound classification and recognition algoritms, because the results of current systems are rather limited. So, why is automatically extracting useful information from many sonic environments not yet successful? The applications of the recent developments require sound source recognition work in complex environments and with flexible tasks. For some of these applications, for example acoustic aggression detection in the public space, the desired information is a single bit: “are there aggressive vocalizations or not?”, for other applications, like in soundscape research, a richer description is required.

Existing techniques for sound recognition are designed to function in closed, specialized domains. Speech recognition and music genre recognition, for example, work under the condition that the input is what they expect; speech from the speaker and the environment the system was trained on, or clean music recordings respectively. The idea that these closed domain techniques will generalize to open environments has so far not materialized. To operate in open environments we need to focus on the constancy and invariants in the signal: the physics that produced it.

In contrast to current “engineered”, specific systems, we aim to develop signal processing techniques that can handle sound in uncontrolled environments. Such environments are outside the range of the problems solved by current techniques, but are the normal environment for humans. These novel techniques are based on the research questions: “How to select sonic evidence that is likely to stem from a single source from a sound signal

recorded in realistic acoustical circumstances?" and "How can the signal, instead of the system design, guide the processing of the signal, towards an optimal rendering of the information in the signal?". In current recognition systems the complete processing of the signal is dictated by the design of the system. This entails that all possible input has to be considered by the designer of the system, which is impossible in open environments. Instead the system should be able to estimate if and to what extend and how the incoming signal should be processed.

The approach taken is based on two observations. First, the fact that many sounds have prominent tone-like and/or pulse-like components. These components correspond to two main sound producing processes, resonance, and impact respectively, and these components are the extremes in localization in frequency (for tones) and time (for pulses). Because of this strong localization the overlap-probability in the time-frequency plane is small for uncorrelated sources, i.e. these sounds are sparse. Second, humans are able to assign all components of a single source to a single representation. I.e. we hear a car or a voice and not the components that constitute the sound. This proces is called object formation and improves the robustness because a group of components is more robust than the components in isolation.

Based on the first observation we developed a efficient method to extract tones and pulses from a time-frequency energy representation. The extraction is based on comparing the energy profile around a time-frequency point with the energy profile around the same point when excited with a pure tone or pulse. The comparison is performed with a sparse matching filter that captures the shape of the excitation in frequency or temporal direction respectively. The resulting measures are called tone-fit and pulse-fit. The tone-fit and pulse-fit

1. are independent of signal level,
2. are complementary when applied to chirps
3. have a strong, well-defined correlation with the local signal-to-noise ratio,
4. are accurate in measuring frequency or time,
5. can separate tones with a relative frequency difference as small as 3%, smaller differences lead to a single component with a informative amplitude modulation,
6. can separate pulses with a time separation equal or greater than the group delay of the filterbank, smaller separations lead to a single component with a informative frequency modulation,
7. correlate with perceptual descriptions of real-world recordings.

Subsets with a high tone-fit or high pulse-fit are extracted. These subsets have a high probability of stemming from a single source due to sparcity. Broadband signals can also produce similar subsets, but these are always small and can be eliminated with a size criterium.

Within each subsets, the energy maxima are strung together in time (for tones) or frequency (for points) to from signal components. If appropriate the tonal components are grouped based on common onset, common frequency modulation and harmonic relation. This grouping improves the robustness of the signal-component further. Apart from tones and pulses an important other class of sounds are broadband sounds. These are extracted by selecting regions that exceed the long-term background for some time and are not classified as either tones or pulses.

During recognition the sound sources feature vectors are extracted for the harmonic groups and broadband events. These feature vectors are classified using using k-nearest neighbor classifiers.

To test the recognition systems, two datasets were created and annotated. The first dataset was recorded on the Amsterdam Amstel train station. Several scene were played by professional actors while the platform was in normal use. The content of these scenes ranged from normal station scenes to aggressive scenes. Recordings were made with eight microphones and three cameras. The dataset was annotated on both common and aggression related sound classes. The second dataset was recorded at several places in the town of Assen (NL) on several different days in different weather conditions by students. The number of sound classes and acoustical environments is much larger than in the first dataset.

Annotations were made by specifying start and stop time and the class for every event. For many classes the start and stop times were found to be ambiguous, due to masking by other sources and personal choices of the annotator. To alleviate part of the tediousness of the annotation work an annotation tool was developed that preselects regions and suggests sound classes based on previous annotations.

Performance on the datasets is measured with the F-measure on frames. This measure is the harmonic mean between recall and precision, and punishes both failure to include frames for a specific class as well as including too many frames. These measures are chosen based on similarities of the identification task with information retrieval. Because the ambiguity in the start and stop times, the overlap of the recognition results with the annotations will not be exact and the F-measure will be lower due to this mismatch, while the recognition result is just as valid. For the Amstel station dataset the F-measure is 0.18 for speech-like classes (“speech”, “singing”, “scream”), partly because of the start and stop time ambiguity, but also because of arbitrary boundaries between the classes. For the “train” and “subway” classes the F-measure is around 0.5. For the Assen dataset the scores are similar, 0.45 for “car”, down to 0.02 for “bird”, for most classes the precision is high and the

recall low. The inter-annotator f-measure for this dataset is 0.46. Though not directly comparable, it indicates that human annotators disagree on the annotations which form the ground truth with the same order as the system (dis)agrees with the ground truth. The signal-driven recognition stage can be complemented with a dynamic network (PhD Thesis M.E. Niessen), which increases the F-measure on average 20% for the Amstel dataset and 33% for the Assen dataset.

The Amstel dataset was also used in an experiment where audio detection results were fused with the results of aggression detection from the video recordings. The combined results showed a higher detection rate (78%) with no more false alarms (16 alarms/hour) than video in isolation (67%, 16 alarms/hour) and audio in isolation (45%, 4 alarms/hour).

Finally, the harmonic extraction was tested on the American-English vowel dataset containing vowel spoken in between “h” and “d”. These vowels were annotated on formant positions and on fundamental frequency. This test showed that the performance of the harmonic complex extraction in clean and moderately noisy situations is good (96%) and only drops significantly around 0 dB signal-to-noise ratio. Performance on the recognition of the vowels is 80% for all speakers classes, with confusion pattern that is not unlike human confusions.

One of the main hurdles in developing systems for automatic sound recognition in everyday situations is the lack of datasets. With the datasets recorded on the Amstel station and in the town of Assen we hope to set the standard for realistic, uncontrolled datasets. These datasets where recorded with as little interference with the environment as possible, while still capturing the events that we wanted to capture.

9.1 Future work

9.1.1 Signal Processing

In the current system, the signal processing is not influenced by the recognition stage nor by the users question. Instead the results of the signal-driven recognition are only reevaluated by the knowledge-driven network. However, it may be beneficial for the signal-driven stage to refine its analysis based on the reevaluated classes. For example after the fan of a laptop is recognized it will be useful to prevent the signal component belonging to the fan from being used in harmonic complexes. On a level closer to the signal, tonal components could be reconnected if they are part of the same harmonic in a harmonic complex, if the presence of energy permits this. However the lower the processing-level the less beneficial the knowledge-driven influence will be.

Although the computational requirements of the signal-processing are

already within the capabilities of modern embedded systems, further improvements are to be made to improve these further. For example, taking the knowledge-driven influence one step further, the system could go in "checking mode" where the system only checks whether expected or earlier detected source are still there.

In this thesis we have focussed on tones and pulses as important types of signals and noise-like signals have only briefly been mentioned (section 4.1.4). However this is an important class of signals, which requires a more statistical, less localized approach. The distributions of the tone-fit and pulse-fit may provide the basis of noise detection and identification, though more broader (in time and frequency) measures may prove to be more effective.

So far all signal processing using the techniques in this thesis have been audio signals. However, all possible signals are bound by the Heisenberg inequality. Therefore the techniques like tone-fit and pulse-fit may be applied to signals of other origin as well.

9.1.2 Annotation

In the current system training and performance measures depend crucially on painstakingly annotated datasets. While these remain necessary for scientific dissemination and detailed performance measurements, for most applications it may be suffice to annotate just the "mid-point" of the event. This kind on annotation can easily be done in realtime for a limited number of known sources. A set of buttons, one per class, would allow the annotator to apply that label to a time instance. As annotators will want to wait to be sure that the "mid-point" has passed an offset could be applied, or video presentation or presence at the scene may allow the annotator to anticipate the "mid-point".

Such an "mid-point" annotation is also more robust then annotating the start and stop times of a sound event as the event is less likely to be masked at its most energetic point in time. For example, the passing of a car is has a clear maximum in the energy, while the start and stop times may be ambiguous due to masking by other sources. Continuous sources, air-conditioning for example, would still require a continuous annotation though as these lack a well-defined "mid-point".

The performance measures, precision, recall and the F-measure are well suited for these "mid-point" annotations. Instead of the per-time version of these measures used in chapters 4 and 7, where length of the annotation/detection had a big influence, the performance would be on event level.

Beside a ground truth, the annotation process can give insight in how people listen to audio recordings. The full-temporal annotation is a very precise and analytical task which may provide the most complete annotation. However this is not a natural listening mode for humans. The "mid-point" annotation allows for less analytical, more natural listening and that

may result in missing sound sources. If time-pressure would be added more sources will go unnoticed. This may provide insights in what people judge to be important sources. This last qualification may also depend elements in the task description. For example, does changing the location mentioned in the task change which sources are deemed important. The same factors should play a role in a automatic sound source recognition system, the system should only invest resources in analyzing the sound to the "start/stop" detail if the task of the system warrants it.

9.1.3 Recognition

The current system classifies unknown sound sources as being "noise", without further analyzing those sounds. A extension would be to use clustering algorithms to group unrecognized sound sources and make them available for annotation. This requires the signal-driven segmentation to work good enough to extract reasonable groups that can be clustered. These techniques could typically be incorporated in the annotation tool introduced in chapter 5.

In combination with other sensors, wind, rain or car detectors a number of classes could be checked automatically. This makes it possible to sidestep the annotation process for these classes. Having multiple sensor(s) (modalities) will also allow the system at large to monitor the functioning of individual sensors.

Besides multiple sensors modalities the recognition at one sensors entails that, when sensible for that class, surroundings sensors could go in "checking mode" for that class. This would allow the system to reduce its overall computational demands and increase it robustness. The human perceptual system uses the same mechanism for exactly the same purpose (Harding et al., 2008).

Finally, the preliminary work done on the recognition of vowels shows encouraging results, but the real challenge in formant detection and vowel recognition is in continuous speech. The research presented in chapter 8 should be expanded to databases of continuous speech like the TIMIT database (Garofolo et al., 1993). Because these databases contain real speech both vowels and consonants should be recognized. For voiced consonants methods similar to those used for vowel can be applied, but for unvoiced consonants techniques for broadband signals should be used. Also the formant trajectories of voiced speech are influenced by adjacent speech sound. Because of this influence the trajectories of the formants of voiced speech may exploited to recognize the unvoiced parts.

10

Samenvatting

A

Gamma-chirp in formulae

The gamma-chirp cochlea-model is implemented as a filterbank with N_{ch} channels or segments:

$$y_{ch}(t) = h_{gc} \otimes x(t) \tag{A.1}$$

where h_{gc} are the filter-coëfficients that make up the gamma-chirp. They are defined as:

$$h_{gc} = at^{N-1}e^{-2\pi bB(f_{ch})t}e^{j(2\pi f_{ch}t + c\log(t))} \tag{A.2}$$

where f_{ch} is the center frequency of the channel, N the order of the gamma-chirp ($N = 4$) and $a = 1$, $b = 0.71$ and $c = -3.7$. The center-frequencies are logarithmical distributed between 60 and 4000 Hz. B is the bandwidth of a filter and is given by the ERB scale Moore and Glasberg (1996):

$$B(f_{ch}) = 24.7 + 0.108 f_{ch} \tag{A.3}$$

To calculate a energy representation from the excitation, the excitation y_{ch} is squared and leaky-integrated. The integration constant τ_{ch} is channel dependent. Finally the result is subsampled and the logarithm is taken to compress the energy into a sensible range:

$$E_{ch}(t) = \int_{t_0}^{t} y_{ch}^2(t-\tau)e^{-\tau/\tau_{ch}}d\tau \tag{A.4}$$

$$\tau_{ch} = \max(\frac{2}{f_{ch}}, 0.005) \tag{A.5}$$

$$E_{dB,ch}(fr) = 10\log_{10}(E_{ch}(fr\ dt_{fr})) \tag{A.6}$$

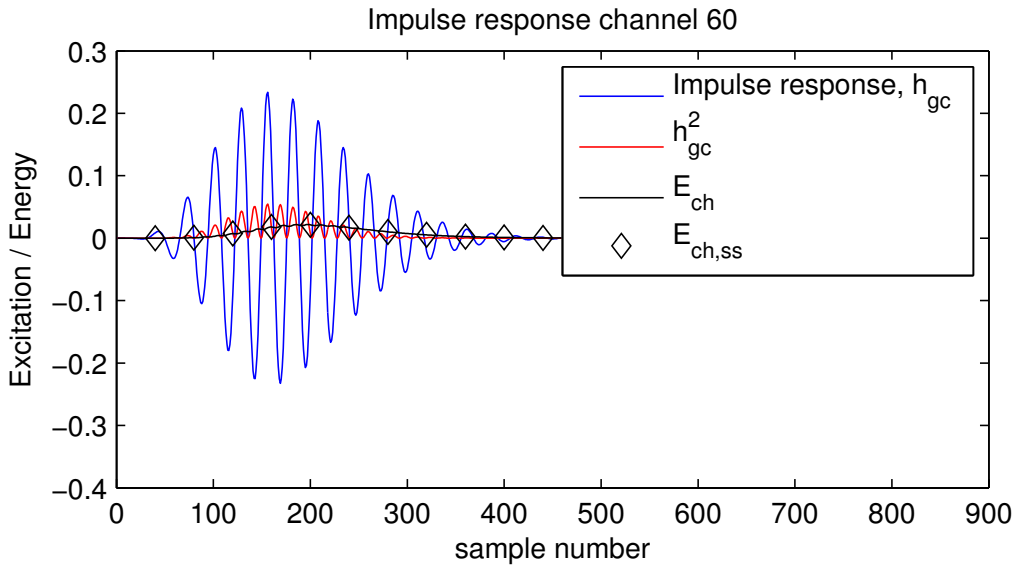


Figure A.1: Several steps in the calculation of the energy representation of a impulse using a gamma-chirp. The step of taking the logarithm is not depicted.

where fr is the frame-number and dt_{fr} the frame-size.

B

List of publications

B.1 Journal papers

- Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C. Andringa. Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters*, 2010.

- Johannes D. Krijnders and Tjeerd C. Andringa. Tone, pulse, and chirp decomposition for environmental sound analysis. *In preparation*, 2010.

B.2 Conference proceedings

- Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C. Andringa. Robust harmonic complex estimation in noise. In *Proceedings of the 19th International Conference on Acoustics*, cas–03–019, 2007.

- Maria E. Niessen, Johannes D. Krijnders, Joep Boers, and Tjeerd C. Andringa. Assessing the reverberation level in speech. In *Proceedings of the 19th International Conference on Acoustics*, cas–03–020, 2007.

- W Zajdel, Johannes D. Krijnders, Tjeerd C. Andringa, and Dariu M. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *Proceedings of 2007 IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 200–205, 2007.

- Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C. Andringa. A grouping approach to harmonic complexes. In *Proceedings of Acoustics '08*, 2008.
- Maria E. Niessen, Ronald A.V. van Elburg, Johannes D. Krijnders, and Tjeerd C. Andringa. A computational model for auditory scene analysis. In *Proceedings of Acoustics '08*, 2008.
- Bea Valkenier, Johannes D. Krijnders, R.A.J. van Elburg, and T.C. Andringa. Robust vowel detection. In *Proceedings of NAG/DAGA 2009*, pages 1306-1309, 2009.
- Maarten van Grootel, Johannes D. Krijnders, and Tjeerd C. Andringa. Research database for everyday listening. In *Proceedings of NAG/DAGA 2009*, pages 996-999, 2009.
- Johannes D. Krijnders, Maria E. Niessen, and Tjeerd C Andringa. Annotating soundscapes. In *Internoise 2009*, , 2009.
- Maria E. Niessen, Johannes D Krijnders, and Tjeerd C Andringa. Understanding a soundscape through its components. In *Proceedings of Euronoise 2009*, EN09_242, 2009.

B

Acknowledgements

References

Jont B Allen. How do humans process and recognize speech? *IEEE Trans. Speech Audio Processing*, 2(4):657–577, Oct 1994. doi: 10.1109/89.326615.

Tjeerd C. Andringa. The texture of natural sounds. In *Proceedings of Acoustics'08*, 2008.

Tjeerd C. Andringa. *Continuity Preserving Signal Processing*. PhD thesis, University of Groningen, 2002.

Tjeerd C. Andringa and Johannes D. Krijnders. Method and system for texture based signal analysis, 2009.

Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J Acoust Soc Am*, 122 (2):881–891, 2007. doi: 10.1121/1.2750160.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Harlow: Addison-Wesley, 1999.

James A Ballas. Common factors in the identification of an assortment of brief everyday sounds. *J Exp Psychol Hum Percept Perform*, 19(2):250–267, 1993.

James A Ballas and James H Howard. Interpreting the language of environmental sounds. *Environment and Behavior*, 19(1):91–114, 1987.

Jon P Barker, Trevor J Cooke, and Daniel P W Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25, Jan 2005. doi: 10.1016/j.specom.2004.05.002.

Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of Uncertainty in Artificial Intelligence*, volume 98, 1998.

Albert S. Bregman. *Auditory Scene Analysis*. The MIT Press, 1990.

Adelbert Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica united with Acustica*, 86(1):117–128, Jan 2000.

Petro Cano. *Content-based audio search: from fingerprinting to semantic audio retrieval*. PhD thesis, Pompeu Fabra University, 2006.

Jean-François Cardoso and Maude Martin. *Independent Component Analysis and Signal Separation*, chapter A Flexible Component Model for Precision ICA, pages 1–8. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007.

Robert P Carlyon, Rhodri Cusack, Jessica M Foxton, and I Robertson. Effects of attention and unilateral neglect on auditory stream segregation. *J Exp Psychol Hum Percept Perform*, 27(1):115–127, Feb 2001. doi: 10.1037/0096-1523.27.1.115.

Edward Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am*, 25(5):975, Jan 1953. doi: 10.1121/1.1907229.

Edward Colin Cherry and W.K Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *J Acoust Soc Am*, 26 (4):554, Jul 1954. doi: 10.1121/1.1907373.

Seungjin Choi, Andrzej Cichocki, HM Park, and SY Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

Selina Chu, Shrikanth Narayanan, and C.-C Jay Kou. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009. doi: 10.1109/TASL.2009.2017438. check references.

Trevor J Cooke. A glimpsing model of speech perception in noise. *J Acoust Soc Am*, Jan 2006.

Trevor J Cooke, Phil Green, Ljubomir Josifovski, and A Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, Jun 2001. doi: 10.1016/S0167-6393(00)00034-0.

COST Action 0804. Soundscape of european cities and landscapes. Technical Report COST 0804, European Cooperation in the field of Scientific and Technical Research, december 2008.

Michael Cowling and Renate Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, 2003. doi: 10.1016/S0167-8655(03)00147-8.

Fabio Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997. doi: 10.1023/A:1006569829653.

Ankur Datta, Mubarak Shah, and Niels Da Vitoria Lobo. Person-on-person violence detection in video data. In *International Conference on Pattern Recognition*, volume 16, pages 433–438, 2002.

Laurent Daudet. A review on techniques for the extraction of transients in musical signals. *Lecture Notes in Computer Science*, Jan 2006.

Michael E Davies and Laurent Daudet. Sparse audio representations using the mclt. *Signal Processing*, Jan 2006.

Bert De Coensel and Dick Botteldooren. Modeling auditory attention focusing in multisource environments. In *Proceedings of Acoustics'08*, 2008.

Febe de Wet, K Weber, Louis Boves, Bert Cranen, Samy Bengio, and Hervé Bourlard. Evaluation of formant-like features on an automatic vowel classification task. *J Acoust Soc Am*, 116:1781, 2004.

John R. Deller Jr., John H. L. Hansen, and John G. Proakis. *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 1999.

Daniele Dubois and Catherine Guastavino. Noise(s) and sound(s): comparing various conceptualizations of acoustic phenomena across languages. In *Proceedings of Acoustics'08*, 2008.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

H Duifhuis, HW Hoogstraten, Sietse M Netten, RJ Diependaal, and W Bialek. *Cochlear Mechanisms: Structure, Function and Models*, chapter Modelling the cochlear partition with coupled Van Der Pol oscilators, pages 395–404. Springer, 1985.

Daniel P W Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication*, 27(3-4):281–298, 1999.

Daniel P.W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.

Antti J Eronen, VT Peltonen, JT Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, G Lorho, and Jyri Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (1):321–329, 2006.

Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proceedings of the 108th Convention of the Audio Engineering Society*, 2005.

Harvey Fletcher. The perception of speech and its relation to telephony. *J Acoust Soc Am*, 22(2):89–151, Jan 1950. doi: 10.1121/1.1906605.

Harvey Fletcher and W A Munson. Loudness, its definition, measurement and calculation. *J Acoust Soc Am*, 5(2):82–108, 1933. doi: 10.1121/1.1915637.

Dennis Gabor. Theory of communication. *Jounal IEE*, 93(26):429–457, 1946.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Palletta, Nancy L Dahlgren, and Victor Zue. Timit acoustic-phonetic continuous speech corpus, 1993.

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

William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993b.

Dariu M. Gavrila. The visual analysis of human movement: A survey. In *Computer Vision and Image Understanding*, volume 73, pages 82–98, 1999.

Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and ...*, Jan 2001.

Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291, 1995. doi: 10.1016/0167-6393(94)00059-J.

Timothy D Griffiths and Jason D Warren. What is an auditory object? *Nature Reviews Neuroscience*, 5:887–892, 2004.

Karlheinz Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser, 1 edition, 2001.

Catherine Guastavino. Categorization of environmental sounds. *Canadian Journal of Experimental Psychology*, 61(1):54–63, 2007.

Brian Gygi, Gary R Kidd, and Charles S Watson. Similarity and categorization of environmental sounds. *Perception and Psychophysics*, 69(6): 839–855, 2007.

Emanuel Anco Peter Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technical University of Eindhoven, 2007.

Sue Harding, Martin Cooke, and Peter König. *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, chapter Auditory Gist Perception: An Alternative to Attentional Selection of Auditory Streams?, pages 399–406. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008.

Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

Simon Haykin. *Blind deconvolution*. Prentice Hall, 1994.

Simon Haykin, editor. *Unsupervised Adaptive Filtering*. Wiley-Interscience, 2000.

Simon Haykin and Zhe Chen. The cocktail party problem. *Neural Computation*, 17:1875–1902, Dec 2005.

James Hillenbrand and Robert A Houde. A narrow band pattern-matching model of vowel perception. *J Acoust Soc Am*, 113(2):1044–1055, Jan 2003. doi: 10.1121/1.1513647.

James Hillenbrand, Laura A Getty, Kimberlee Wheeler, and Michael J Clark. Acoustic characteristics of american english vowels. *J Acoust Soc Am*, 95 (5):2875–2875, 1994. doi: 10.1121/1.409456.

Rolf Hut, Marinus Boone, and Andries Gisolf. Cochlear modeling as time-frequency analysis tool. *Acustica united with Acustica*, 92(4):629–636, 2006.

Toshio Irino and Roy D Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *J Acoust Soc Am*, 101(1):412–419, Jan 1997. doi: 10.1121/1.417975.

Katherine N Irvine, Patrick Devine-Wright, Sarah R Payne, Richard A Fuller, Birgit Painter, and Kevin J Gaston. Green space, soundscape and urban sustainability: an interdisciplinary, empirical study. *Local Environment*, 14(2):155–172, 2009. doi: 10.1080/13549830802522061.

Jean-Claude Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *J Acoust Soc Am*, 93(1):510–524, 1993. doi: 10.1121/1.40563.

Walter Kellerman. *Handbook of Signal Processing in Acoustics*, chapter Beamforming for Speech and Audio Signals, pages 691–702. Springer, 2009.

Johannes D Krijnders and Tjeerd C Andringa. Demonstration of online auditory scene analysis. In *Proceedings of Belgian Netherlands Artificial Intelligence Conference*, 2008.

Johannes D Krijnders and Tjeerd C Andringa. Tone, pulse, and chirp decomposition for environmental sound analysis. *IEEE Trans. Speech Audio Processing*, 2009. Page 2 in of.

Johannes D Krijnders, Maria E Niessen, and Tjeerd C Andringa. Robust harmonic complex estimation in noise. In *Proceedings of the 19th International Conference on Acoustics*, pages cas–03–019, 2007.

Johannes D Krijnders, Maria E Niessen, and Tjeerd C Andringa. A grouping approach to harmonic complexes. In *Proceedings of Acoustics '08*, 2008.

Johannes D Krijnders, Maria E Niessen, and Tjeerd C Andringa. Annotating soundscapes. *Internoise 2009*, 2009.

Johannes D Krijnders, Maria E Niessen, and Tjeerd C Andringa. Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters*, 2010. doi: 10.1016/j.patrec.2009.11.004.

Richard P Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, Jul 1997. doi: 10.1016/S0167-6393(97)00021-6.

Ruth Y Litovsky, H Steven Colburn, William A Yost, and Sandra J Guzman. The precedence effect. *J Acoust Soc Am*, 106(4):1633–1654, 1999. doi: 10.1121/1.427914.

Leendert Van Maanen, Hedderik van Rijn, Maarten Van Grootel, Stephanie Kemna, Martin Klomp, and E Scholtens. Personal publication assistant: Abstract recommendation by a cognitive model. *Cognitive Systems Research*, 11:120–129, march 2008.

Michael E Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6):830–864, 2000.

Sylvain Marchand and Philippe Depalle. Generalization of the derivative analysis method to non-stationary sinusoidal modeling. In *Proceedings of the 11th Int. Conference on Digital Audio Effects*, 2008.

Ranier Martin. Noise power spectral density estimation based on optimal smoothingand minimum statistics. *Speech and Audio Processing*, 9(5): 504–512, Jul 2001. doi: 10.1109/89.928915.

Robert J McAulay and Thomas F Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Speech Audio Processing*, 34 (4):744–754, Aug 1986.

James L McClelland and David E Rumelhart. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, 88(5):375–407, 1981.

Erin McKean, editor. *New Oxford American Dictionary*. Oxford University Press, 2nd edition.

Brian CJ Moore and B Glasberg. A revision of zwicker's loudness model. *Acustica united with Acustica*, 82(2):335–245, Mar 1996.

Kamran Mustafa and Ian C Bruce. Robust formant tracking for continuous speech with speaker variability. *IEEE Trans. Speech Audio Processing*, 14 (2):435—444, Mar 2006. doi: 10.1109/TSA.2005.855840.

Anna K Nábělek and Pauline K Robinson. Monaural and binaural speech perception in reverberation for listeners of various ages. *J Acoust Soc Am*, 71(5):1242, 1982.

Maria E. Niessen. *Context-based sound event recognition*. PhD thesis, University of Groningen, 2010.

Maria E Niessen, Johannes D Krijnders, Joep Boers, and Tjeerd C Andringa. Assessing the reverberation level in speech. In *Proceedings of the 19th International Conference on Acoustics*, pages cas–03–020, 2007.

Maria E Niessen, Gert Kootstra, Sjoerd De Jong, and Tjeerd C Andringa. Expectancy-based robot localization through context evaluation. *International Conference on Artificial Intelligence*, pages 371–377, 2009a.

Maria E Niessen, Leendert Van Maanen, and Tjeerd C Andringa. Disambiguating sounds through context. *International Journal on Semantic Computing*, 2(3):327–341, 2009b.

Douglas O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, Oct 2008. doi: 10.1016/j.patcog.2008.05.008.

ITU-T P.56. Objective measurement of active speech level, March 1993.

Alex Pentland. Smart rooms. *Scientific American*, 274(4):54–62, Jan 1996.

Victor M A Peutz. Articulation loss of consonants as a criterion for speech transmission in a room. *Journal of Audio Engineering Society*, 19(11): 23–27, 1971.

M.Ross Quillian. *Semantic Information Processing*, chapter Semantic memory, pages 216–270. MIT Press, 1968.

Sundarrajan Rangachari and Philipos C Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48 (2):220–231, Feb 2006. doi: 10.1016/j.specom.2005.08.005.

Axel Röbel. Frequency slope estimation and its application for non-stationary sinusoidal parameter estimation. In *Proceedings of the 10th Int. Conference on Digital Audio Effects*, 2007.

Nicoleta Roman, Soundararajan Srinivasan, and DeLiang Wang. Binaural segregation in multisource reverberant environments. *J Acoust Soc Am*, 120:4040–4051, 2006.

Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigne, and Shigeki Sagayama. Single and multiple f_0 contour estimation through parametric spectrogram modeling of speech in noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1135–1145, 2007.

ANSI S1.4. Standard for sound level meters, 2001.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

R. Murray Schafer. *The soundscape, Our sonic environment and the tuning of the world*. Destiny books, 1977.

Klaus R Scherer. Vocal affect expression : A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, 1986. doi: 10.1037/0033-2909.99.2.143.

Valeriy Shafiro and Brian Gygi. How to select stimuli for environmental sound research and where to find them. *Behavior Research Methods*, 36 (4):590–598, 2004.

Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

Barbara G Shinn-Cunningham. Object-based auditory and visual attention. *TRENDS in Cognitive Sciences*, 12(5):182–186, May 2008. doi: 10.1016/j.tics.2008.02.003.

Harry L. Van Trees. *Optimum Array Processing*. Wiley-Interscience, 1st edition, 2002.

Barry Truax. *Acoustic Communication*. Greenwood Publishing Group, 2nd edition, 2001. ISBN 9781567505368.

Bea Valkenier, Johannes D Krijnders, Ronald AJ Van Elburg, and Tjeerd C Andringa. Robust vowel detection. *NAG/DAGA 2009*, 2009.

Fredinand van der Heijden. *Image Based Measurement Systems*. Wiley, 1994.

Maarten van Grootel, Tjeerd C. Andringa, and Johannes D. Krijnders. Dares-g1: Database of annotated real-world everyday sounds. In *Proceedings of NAG/DAGA 2009*, 2009.

Peter W.J. van Hengel and Tjeerd C Andringa. Verbal aggression detection in complex social environments. In *Proceedings of 2007 IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 15–20, 2007. doi: 10.1109/AVSS.2007.4425279.

Leo Paulus Antonie Servatius van Noorden. *Temporal Coherence in the Perception of Tone Sequences*. PhD thesis, Technical University of Eindhoven, 1975.

C. J. van Rijsbergen. *Information Retrieval*, pages pp.112–140. Butterworths, 2 edition, 1979.

Nancy J. Vanderveer. *Ecological acoustics: Human perception of environmental sounds.* PhD thesis, Cornell University, 1979.

Julio Vargas and Steve McLaughlin. Cascade prediction filters with adaptive zeros to track the time-varying resonances of the .... *IEEE Trans. Speech Audio Processing*, 16(1):1–7, Jan 2008. doi: 10.1109/TASL.2007.907573.

DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis*. IEEE Press / Wiley-Interscience, 2006.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2 edition, 2005.

Qin Yan, Saeed Vaseghi, Esfandiar Zavarehei, Ben Milner, Jonathan Darch, Paul White, and Ioannis Andrianakis. Formant tracking linear prediction model using hmms and kalman filters for noisy speech processing. *Computer Speech & Language*, 21(3):543–561, 2006.

Wojtek Zajdel, Johannes D. Krijnders, Tjeerd C. Andringa, and Dariu M. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *Proceedings of 2007 IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 200–205, 2007. doi: 10.1109/AVSS.2007.4425310.

Miroslav Zivanovic, Axel Röbel, and Xavier Rodet. Adaptive threshold determination for spectral peak classification. *Computer Music Journal*, 32(2): 57–67, 2008.

Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

Zoran Zivkovic and Ben Kröse. An em-like algorithm for color-histogram-based object tracking. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.