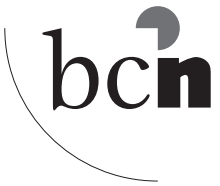
Context-Based Sound Event Recognition

This research was generously supported by:

SenterNovem (Dutch Companion project grant number ISO53013)
School of Behavioral and Cognitive Neurosciences
Sony Computer Science Laboratory Paris



Printed by: Gildeprint, Enschede
Cover photo: © 2010 INCAS^3 / Deborah Roffel
ISBN: 978-90-367-4542-0

RIJKSUNIVERSITEIT GRONINGEN

Context-Based Sound Event Recognition

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
vrijdag 22 oktober 2010
om 16.15 uur

door

Maria Elisabeth Niessen

geboren op 5 oktober 1980
te Apeldoorn

Promotor:	Prof. dr. L. R. B. Schomaker
Copromotor:	Dr. T. C. Andringa
Beoordelingscommissie:	D. Dubois, Directrice de recherche CNRS Prof. dr. H. J. van den Herik Prof. dr. B. G. Shinn-Cunningham

Contents

APPLICATIONS

1

General introduction

A simple game, named "guess the sound", has been successful on the radio for several decades. Radio listeners have to guess the source of a short audio fragment, such as a match box being opened, shuffling playing cards, opening an umbrella, and so forth. Usually a prize can be won by giving the correct answer. For example, on a Dutch radio channel the value of the prize that can be won increases as the number of failures at guessing the audio fragment increases. Furthermore, the radio channel gives hints that should make guessing the sound easier, such as "you hold it in your hands" in case of shuffling playing cards. The difficulty of the game can be demonstrated by an analogy in the visual domain: try to recognize what the object that is depicted in Figure 1.1 represents. Why can money be won with a task that people perform effortlessly in their everyday life? The answer is related to the hints that the radio channel gives out: people use the context in which a sound is heard to identify it. When they hear sounds (or see objects) in their everyday life, they use their knowledge of the environmental context to infer which events they are likely to hear, and to discard interpretations that are unlikely given the context. Moreover, people do not only hear, they also see, smell, and feel. Therefore, they are normally not as clueless as the radio listener who tries to guess the sound.

Current automatic sound recognition systems are faced with the same problem as the radio listener. The system is presented with a short audio fragment, and has to recognize the source of the sound. The task is usually even more complicated, because the input cannot be controlled if the *audio signal*[1] is recorded in a real-world environment. As a consequence, the audio signal can comprise *sound* produced by multiple sources. In this thesis we will demonstrate that the task of the automatic sound recognition system can be alleviated in the same manner as for the radio listener, namely by giving it knowledge about the context of the sound. To accomplish this goal we will first explore the problem of recognizing sound events in a real-world environment by reviewing strategies of both automatic systems and human listeners. Subsequently, we introduce and test a model that incorporates context into an automatic sound recognition system.

[1] Whenever we introduce a term that is important in this thesis, it is printed in italic. The definitions of these terms can be found in the glossary (appendix A).

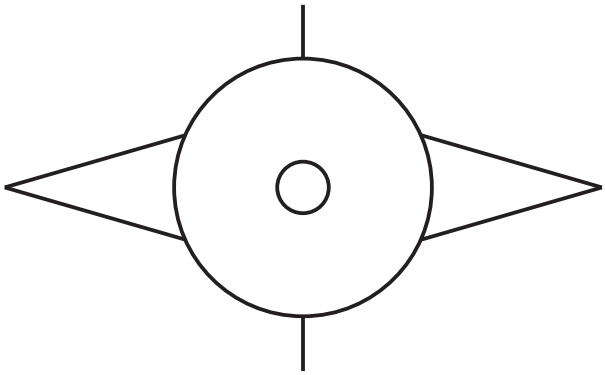


Figure 1.1: A visual analogy of "guess the sound". The image is a two-dimensional schematic representation of a visual landscape, like the audio sample is stereo playback of a sonic environment. Because the context information is removed, it is difficult to guess what the image represents. Interpretation is easier when hints are given about the context of the image. For example, this image is viewed from the top, and water is surrounding the depicted object.

1.1 AUTOMATIC SOUND RECOGNITION

The research domain of automatic sound recognition aims at describing an audio signal in terms of the sound events or sources that compose a sonic environment. It has important (future) applications in fields as diverse as monitoring sonic environments, robotics, security systems, content-based indexing of multimedia files, and human-machine interfaces. Most sound recognition research is aimed at improving one of these application domains, such as speech recognition (O'Shaughnessy, 2008) or music genre classification (Aucouturier and Pachet, 2003). The methods in these application domains have proven successful in problems with a single, known type of sound, and even in recognizing isolated environmental sounds, such as footsteps or jangling keys (Cowling and Sitte, 2003). However, these methods have some attributes that make them less suitable for automatic sound event recognition in real-world environments.

To recognize a sound event implies that it is already known to the receiver. Therefore, a representation of the sound event needs to be stored, so it can be retrieved when an instance of the sound event is encountered. In automatic sound recognition systems this representation is typically stored as a set of temporally ordered features that describe the whole spectrum of each time frame (short time interval) in the signal—a spectrum is a representation of the energy contribution for all frequencies in one frame of an audio signal (Figure 1.2). Intervals of an audio signal or pre-selected samples are classified based on these features. However, in an audio signal of a real-world environment these intervals or samples do not necessarily correspond to (part of) a single event, because the amount of sound events contributing to the audio signal is not controlled. Multiple sound events can co-occur, at other moments no sound events occur at all. Furthermore, the audio signal can be masked or distorted by transmission effects such as background sounds and reverberation. As a consequence, a system that has to recognize sound events in a real-world environment cannot rely on the assumption that the input consists of a single, known, and undistorted signal type, as the methods used on speech and music often can.

A system for real-world sound recognition needs to *segregate* possible sound events from a background before it can recognize the events. Segregation of a sound event means that its constituent components in the time-frequency plane—a representation that shows the temporal development of frequency components in an audio signal (see appendix B)—are selected and grouped. As a result, correctly segregated events can be analyzed as individual elements of a sonic environment. Figure 1.3 shows a schematic example of selection and grouping in a time-frequency plane. Even if sound events are segregated and undistorted, different events can share similar *audio pattern*s (combination of components) while they convey a distinct meaning, such as screaming and laughing. In other words, the event that produces the sound can be ambiguous, similar to the fragments in "guess the sound". In conclusion, we need two constituents to make an advance in real-world sound event recognition. First, the segregation of audio patterns from a signal should provide hypotheses about the sound events producing these patterns. Second, we need a model that interprets these hypotheses based on contextual knowledge, and disambiguates sound events if necessary. The second task is the focus of this thesis.

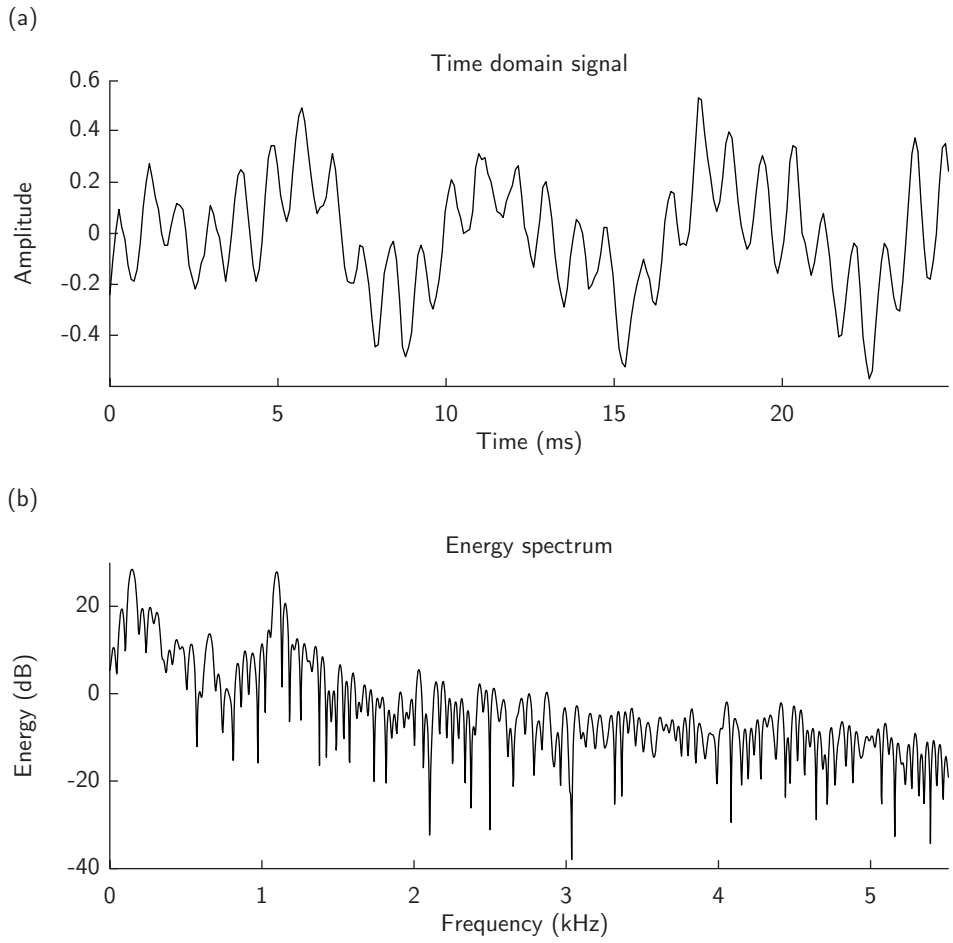


Figure 1.2: One time frame of 25 milliseconds (a) and the energy spectrum of this frame (b) from an audio signal of the sound of a siren with rain and thunder in the background. The siren produces a tonal sound with a frequency around 1.1 kHz, which can be calculated from the fast zero crossing in the time domain signal, and it can be seen as a peak in the energy spectrum. Furthermore, a slower wave can be seen in the time domain signal, which is also visible as a peak around 140 Hz in the spectrum. This frequency component is caused by the thunder. The rain is a noise-like sound. Hence, its spectrum covers a broad frequency range. A feature that describes the whole spectrum includes information about all sound events that are present in the audio signal (the siren, thunder and rain).

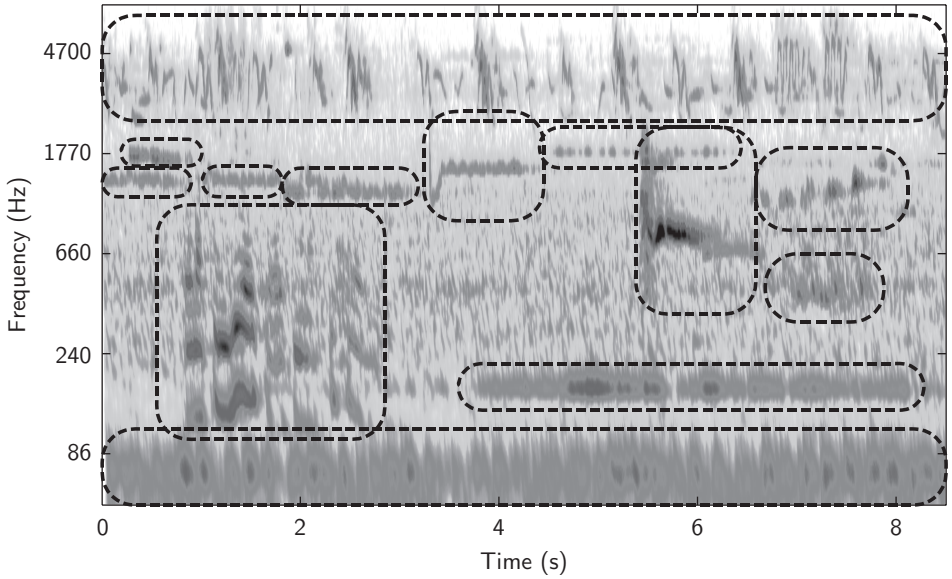


Figure 1.3: The time-frequency plane of a recording at a square in a town (with sound events like speech, birds, and a plane), computed with a gammachirp filter bank (Irino and Patterson, 1997). The gray-scale indicates the energy in decibels (dB): darker gray corresponds to more energy. The frequency axis is logarithmic. The dashed boxes indicate possible ways to select and group events in the plane based on their local properties. For example, between approximately 1 and 3 seconds speech can be seen, which can be grouped because its components are harmonic.

1.2 HUMAN SOUND PERCEPTION

While present-day automatic sound recognition is often designed for specific tasks or specific environments, human sound perception functions for all sounds and is robust to many different environmental conditions that influence the audio signal. People can recognize a driving car, whether they hear it on a road or on gravel, in the rain or in a tunnel. Even when people have not heard a sound event before, they are able to hypothesize as to the source of the event. This ability does not rely only on auditory processing, but includes many cognitive functions as well. Depending on factors like their goals, expectations, memory, preferences, and current situation, people perceive the sound events in their environment differently.

For example, in a familiar environment people do not have to identify common sound events, because they are expected, and do not provide any new information (Grossberg, 1980).

One important factor that allows people to hear in unconstrained environments is their knowledge of the context, which helps them to form predictions and guide their perception of the environment (Bar, 2007). Events or objects in the real-world usually do not occur in isolation, but are related to other events and can be heard in particular environments (Oliva and Torralba, 2007). Therefore, the meaning of a sound event pertains to the associations that people have to other events and environments. Especially when an audio signal is unreliable or can be interpreted in multiple ways, associations help them to recognize an event (cf. Figure 1.1[1]), a phenomenon that is primarily investigated in visual perception (Bar, 2004). For example, when parts of continuous speech are replaced with noise, people can still perceive the speech as being continuous. Furthermore, their interpretation of the distorted part of the speech depends on the meaning of the surrounding speech (phonemic restoration, Warren, 1970; Samuel, 1996).

1.3 SOUND SOURCES, EVENTS, AND PERCEPTS

In the previous sections we have introduced how automatic systems and people recognize sound events. However, we have not yet clarified the relation between a source, an event, and the perception of an event. In contrast to vision, audition is by default not static; that is, something in the world has to happen or change to produce sound. Furthermore, a potential *sound source* has to be involved in some event to produce sound, and it can often produce multiple types of *sound event*s. For example, a car can be parked, producing no sound, someone can accelerate it, producing a sound event that is caused by a process in the engine, or someone can stop the car, producing a different sound event that is caused by a different process. Sound event recognition describes the task of recognizing events, which are caused by a physical action involving a source.

Human perception of sound events is explained by several theories. The ecological approach to sound perception adopts the term 'everyday listening' to refer to listening to sound events in everyday life (as opposed to 'musical listening',

[1] A mexican in a canoe.

Gaver, 1993). Its focus is on the invariant (constant in different situations) perception of the physics of an event. For example, a large group of studies has focused on the ability of people to hear some physical property in the sound produced by an object, such as the perception of object size (Carello *et al.*, 1998; Kunkler-Peck and Turvey, 2000), or the ability to distinguish between bouncing or breaking objects (Warren and Verbrugge, 1984). However, ecological psychology is less concerned with the functional process of recognition involving the role of memory in perception. Because this aspect is essential for modeling real-world sound event recognition, our focus is more toward the information processing approach developed in cognitive psychology than on ecological psychology.

The information processing approach analyzes cognition and perception by abstract stages in the processing of a task (Anderson, 2005). Cognitive psychology is not concerned with the function of the brain. Hence, these abstract stages do not necessarily correspond to the processing stages in the brain. Theories about cognition are usually investigated with an experimental paradigm to confirm (or falsify) the hypothesized stages. This experimental approach was first empirically tested by Sternberg (1966) in a memory decision task. In a similar way, auditory perception can be analyzed as a succession of conceptual processing stages, from sensory transduction, via auditory grouping and categorization, to recognition (McAdams, 1993). A schematic overview of human perception of a sound event is depicted in Figure 1.4. When people perceive an object or event in the world, they match the grouped percept to a prototype or exemplar of a category (Rosch, 1975; Dubois, 2000)—categories are not represented in memory by membership conditions, but rather by the attributes of a prototype (Reed, 1972; Smith and Minda, 2000) or exemplar (Medin and Schaffer, 1978; Nosofsky and Zaki, 2002).[1] Furthermore, this prototype triggers associations in memory, which provide predictions about properties of the environment, for example, what may be perceived next.

[1] For example, when one tries to think of list of properties shared by all chairs, it will be quite impossible to come up with even one. For every property an exception can be found that would still be conceived of as a chair. Yet, when people see an object, they have no difficulties in determining whether or not it is a chair.

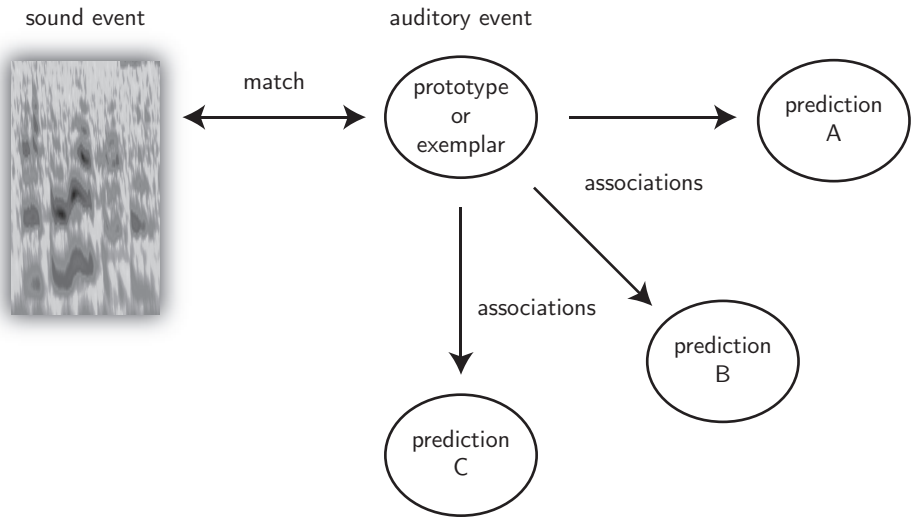


Figure 1.4: A schematic overview of human perception of a sound event. The perceived event is matched to a prototype or exemplar of an auditory event category, which triggers associations in memory. These associations provide predictions about properties of the environment (Bar, 2007).

1.4 OVERVIEW

The thesis is divided in two parts. In the first part we give fundamental background on recognizing sound events. Furthermore, we introduce a method that is based on a model of human memory to improve sound event recognition with knowledge of the context. In the second part we demonstrate the improved performance of two applications that integrate *signal-driven* methods with the context model proposed in the first part.

Real-world environments pose additional demands on sound event recognition over controlled conditions, such as an input of isolated sounds. Chapter 2 discusses how these demands can be managed by an automatic system for sound event recognition. In chapter 3 we review the research on human perception of sound events, with a focus on studies in cognitive science. Furthermore, we present an experiment to demonstrate how context can facilitate sound event recognition. This facilitatory effect is known in visual perception, but hardly investigated in auditory perception. Finally, in chapter 4, we introduce a model that incorporates context

into an automatic sound event recognition system, based on the findings in the previous two chapters.

Although sound recognition in real-world environments has been used to distinguish between different types of sonic environments, such as parks and roads (Aucouturier *et al.*, 2007), automatic recognition of the sound events that constitute a sonic environment is a new area of research with important applications that require a different approach. In chapter 5 we present two experiments that show the possibility to automatically recognize sound events in an unconstrained environment with the combination of techniques for sound event segregation (Krijnders, 2010) and the context model. Moreover, the context model is not limited to audio input, but can be applied to input signals from other modalities as well, as is demonstrated in chapter 6. In this chapter we apply the context model to improve robot localization by disambiguating visual observations of a mobile robot.

2

Challenges of sound recognition in the real world

The content of section 2.4 has been published as Niessen, M. E., Krijnders, J. D., Boers, J., & Andringa, T. C. (2007). Assessing the reverberation level in speech. In Proceedings of the 19th International Congress on Acoustics.

Systems that operate in a real-world environment have to process ambiguous and noisy input. Current techniques for sound recognition are mostly designed for specific applications, such as automatic speech recognition. Therefore, they can rely on assumptions about the audio signal, such as the signal being undistorted and of a known type. However, these assumptions cannot usually be met in a real-world environment. Therefore, real-world sound event recognition requires methods to segregate individual sound events from a globally sounding environment. Furthermore, a system that operates in an uncontrolled environment needs to handle transmission effects. To be able to function reliably, the system should be able to adapt to a variety of situations. However, it is not necessary to solve the problems of real-world environments only with signal-driven methods.

2.1 INTRODUCTION

Systems that operate in a real-world environment are confronted with additional challenges compared to systems that operate in a simulated or controlled environment. They have to process an abundance of information, of which not everything is necessarily relevant (Van Hengel and Andringa, 2007). Moreover, sensory information is likely to be ambiguous or noisy. In section 2.2 we evaluate whether classification methods used in automatic sound recognition are suitable to recognize sound events in real-world environments. In section 2.3 we discuss the problem of separating sound events from the background in order to recognize them. The recognition of sound events in real-world environments is further complicated by transmission effects, such as reverberation and concurrent sources. Possible approaches to deal with transmission effects are discussed in section 2.4. Finally, in section 2.5, we conclude with the implications of these challenges for robust real-world sound event recognition.

2.2 STATE OF THE ART IN AUTOMATIC SOUND RECOGNITION

Automatic sound recognition has important (future) applications in fields as diverse as environmental noise monitoring, robotics, security systems, content-based indexing of multimedia files, and human-machine interfaces. Most sound recognition research is aimed at improving one of these application domains, such as speech recognition or music genre classification. Typically, the techniques used in these applications classify a sound sample as one class of a closed set of learned classes, of which the descriptors match the descriptors of the sample best. Because these techniques are applied to a known type of sound, they can apply specialized features to describe the sound (Davis and Mermelstein, 1980). For example, within music genre classification (Tzanetakis and Cook, 2002; Aucouturier and Pachet, 2003) and speech recognition (O'Shaughnessy, 2008), spectral-based features such as Mel frequency cepstral coefficients (MFCCs) capture important information of harmonic sounds, but are not very robust to noise (O'Shaughnessy, 2008). In addition, methods for automatic speech recognition rely on a temporal ordering of the signal, which is exploited by searching for the most probable sequence of hidden Markov models (HMMs, Juang and Rabiner, 1991).

Cowling and Sitte (2003) tested a selection of feature extraction techniques and classification methods for speech and music signals on isolated environmental sounds. All methods perform better on speech sounds than on environmental sound events, because environmental sounds exhibit more diverse acoustic properties than speech. However, the best results (70% classification rate for eight classes) suggest that some classification techniques can be effectively applied to the recognition of isolated environmental sounds. Similar to applications in speech and music, the input is controlled so that the methods can classify it. In other words, the input has content that belongs to a single class that is a member of a limited set of known classes, although the type of sound is different from and more diverse than speech and music. In contrast, Defréville *et al.* (2006) applied multiple features to classify samples from continuous real-world recordings. Their results varied from 72% to 99% classification rate per class for six classes.

Another method for sound analysis, the bag-of-frames (BOF) method, has been shown to be able to identify auditory scenes from real-world recordings, such as the street where a recording has been made (Aucouturier *et al.*, 2007). However, the BOF method is not designed to represent details about individual events 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 recognition of sound events. For example, the estimated location of a recording, such as a street or a park, can be useful to infer probable sound events that may occur.

Although the combination of whole-spectrum feature extraction and classification has proven useful in problems with a single known signal type, and even in environmental sound recognition, these methods have some attributes that make them less suitable for automatic sound event recognition in real-world environments. Sound classification methods either classify pre-selected samples with one type of sound, or segment a signal into different intervals based on its acoustic properties. However, these intervals do not necessarily correspond to events. A system for real-world sound event recognition needs to segregate the individual events from the background before it can classify the events. Sound events in a real-world environment often co-occur, at other moments no (interesting or recognizable) events may take place. In other words, we do not want to know of certain intervals what type of sound they are, but which events are present at what time

in a continuous audio signal. In variable real-world environments, a method for sound event recognition cannot assume known and uniform input, as the methods used on speech and music can.

2.3 SOUND EVENT SEGREGATION

To be able to recognize individual sound events in a globally sounding environment (see Figure 1.3), the *audio component*s that constitute an event need to be separated and grouped from the background. We refer to the joint process of separating components and grouping them as segregation. Practical applications of sound recognition, such as automatic speech recognition, have advanced research in sound segregation methods. Strategies for segregation include spatial separation of audio input from microphone arrays (beamforming, Brandstein and Ward, 2001), and independent component analysis (ICA, Jutten and Herault, 1991). Although these methods are successful in the applications they are designed for, they rely on assumptions that cannot be met in all situations. For example, beamforming approaches are challenged by moving targets and by multiple targets close to each other. ICA assumes that the sound events in a mixture are statistically independent. Consequently, these methods are not suitable as a general approach to sound event segregation. To address the issues of application-driven methods, a field of research has emerged that is inspired by human perceptual mechanisms, called computational auditory scene analysis (CASA, Wang and Brown, 2006).

The common aim of CASA studies is to infer properties of individual sound sources from an auditory scene based on either a mono or a stereo recording. CASA research is primarily aimed at analyzing speech and music sounds.[1] Therefore, the target sound is defined by the problem, namely segregating speech or music sounds, either from each other or from background sounds. However, two important aspects of human auditory scene analysis (ASA) are underrated by these approaches. First, human ASA is not limited to segregating harmonic sounds. In real-world environments people are confronted with many other sound classes as well, such as machine, traffic, and nature sounds. Second, whether a sound is tar-

[1] The topic query "computational auditory scene analysis" gives 222 results in ISI Web of Knowledge and about 1850 in Google Scholar. Excluding the results that also contain one or more of the terms "speech", "music", "voice", "harmonic", and "pitch" leaves only 38 and 86 results respectively.

get or background sound does not depend on the audio signal. Instead, it depends on factors such as the goal and expectancy of the listener (or application): "... all sound sources are potential signals and noises. Whether or not a sound from a particular source is signal (wanted) or noise (unwanted) depends on non-auditory events" (Yost, 1991, p. 16). As a result, a method for general purpose sound event segregation cannot define in advance what the target sound is, and what the background noise. Instead, it should provide hypotheses of what can be inferred from the audio signal based on physical knowledge.

2.4 TRANSMISSION EFFECTS

In addition to segregating hypotheses about sound events in the audio signal, a system that analyzes sound in a real-world environment has to deal with transmission effects such as concurrent sources and reverberation. Reverberation leads to a mixing of the target sound with a time delayed version of itself. Therefore, both the effects of concurrent sources and reverberation are similar to the problem of sound event segregation, discussed in the previous section. However, the effects of reverberation can be reduced with knowledge of the source and the sonic environment, which facilitates sound event segregation.

The reduction of effects of reverberation have mostly been studied for speech sounds, since the performance of automatic speech recognition (ASR) systems is affected severely by distortions in the signal (Figure 2.1 shows an example of a reverberant speech signal compared to the clean signal). Classifiers in ASR systems, such as hidden Markov models (HMMs), use features that describe the whole spectrum (see section 2.2). When a signal is distorted by a delayed version of itself, irregular frequency dependency patterns of constructive and destructive interference cause rapid energy fluctuations in the frequency content. Hence, the signal descriptors of a reverberant audio signal will deviate from the descriptors in clean training conditions.

One solution to deal with reverberant environments is to reduce the mismatch between the training conditions and the operating conditions. For example, if HMMs are trained on reverberant speech, ASR will perform better in similar reverberant operating conditions than if the HMMs are trained on clean speech (Matassoni *et al.*, 2000). However, the effects of room acoustics vary greatly for different

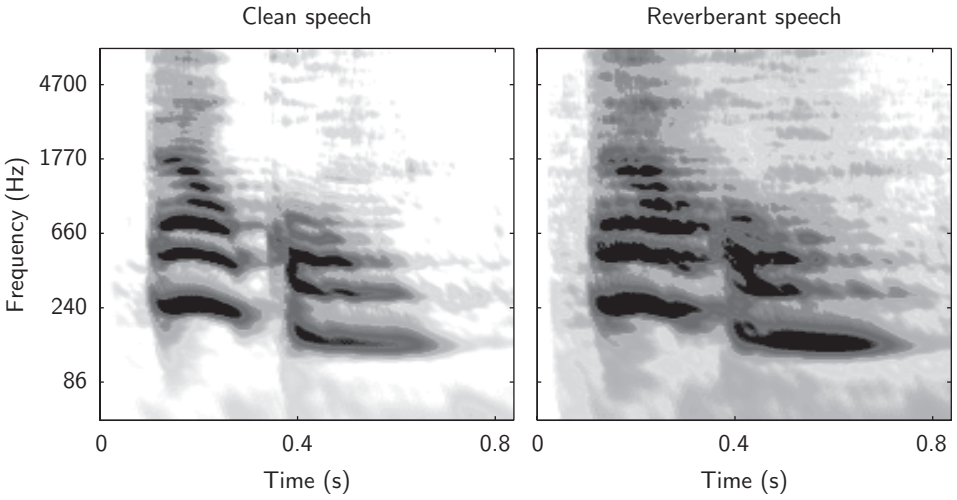


Figure 2.1: The time-frequency plane (computed with a gammachirp filter bank) of a clean speech signal on the left, and of the same speech in a reverberant environment (reverberation time is 320 milliseconds) on the right. The gray-scale indicates the energy in decibels (dB): darker gray corresponds to more energy. The frequency axis is logarithmic. When harmonic components are stationary, reflections may result in an increased energy of the component, and a longer duration. Reflections of variable components cause irregular distortions in the signal.

environments. Different parameters, such as the size of the room, the material on the floor and walls, and the temperature, influence the acoustic characteristics (Kuttruff, 1979). Therefore, this type of ASR system requires training data that match the characteristics of the operating environment. Couvreur and Couvreur (2004) propose a method where acoustic models are trained on speech under different, simulated reverberant conditions. During operation of the ASR system the model that matches the operating conditions best is selected. They show an improved performance on simulated reverberated speech compared to an ASR system trained on clean speech. However, the improvement on realistic data is not as high as on the data with simulated reverberation, because of the discrepancy between real reverberant and simulated reverberant speech.

Another approach to resolve the discrepancy between the training data and reverberant data is to recover the clean speech from the reverberant signal, instead of

adjusting the training data to the operating conditions. An inverse filter is applied to the reverberant signal to remove the distortion caused by reflections, based on the estimated impulse responses of the environment. However, inverse filtering relies on known and stable acoustic characteristics of the environment. As a consequence, methods based on inverse filtering are not robust to changes in the environment, such as the position of the source or the microphone (Radlovic *et al.*, 2000).

Other methods have been designed that are more robust to reverberation in an unknown, but stable, environment. For example, cepstral mean subtraction (Kinoshita *et al.*, 2009) can handle early reflections in ASR. Additionally, late reflections can be suppressed through spectral subtraction (Wu and Wang, 2006; Kinoshita *et al.*, 2009). When the environmental conditions are sufficiently stable, these methods improve the results of ASR.

To test whether it is possible to assess the reverberation level of a monaural audio signal in an unknown and possibly variable environment, we developed a method to classify the reverberation level of speech signals. Reverberation causes an increase in the variation of the energy and frequency of harmonics in speech. Hence, features that capture this variation can be useful to estimate the reverberation level of a signal without a priori knowledge of the environment. We designed and measured such features on speech samples with different levels of reverberation. Clean speech was artificially reverberated to be able to test a controlled set of conditions.

2.4.1 Methods

A common measure for the level of reverberation is the reverberation time T_{60}. The reverberation time is defined as the time for the sound energy level to decay 60 dB after the excitation has ended. We computed nine different levels of reverberation using the Eyring-Norris equation (Eyring, 1930; Norris and Andree, 1929; Kuttruff, 1979):

$$T_{60} = \frac{0.161V}{4mV - S\ln(1-\bar{a})}, \tag{2.1}$$

where V is the room volume in cubic meters, m is a vector with air absorption coefficients for the frequency bands, S is the total surface area, and $\bar{a}$ is the mean wall absorption coefficient. We assumed a fixed room size of 10 by 12 by 3.5 meters

and a constant temperature and humidity of 20 °C and 60% respectively. Hence, the mean wall absorption coefficient was the only variable parameter. Values were assigned to this parameter such that we had a collection of nine reverberation levels ranging from no reverberation to a reverberation time of approximately 1.6 seconds.

The reverberation level can also be expressed by the reverberation radius or distance (Kuttruff, 1979). The reverberation radius is the distance from the speaker or microphone to the sound source for which the energy contribution of the direct sound and the reflected energy are equal. A more reverberant environment coincides with a smaller reverberation radius. Naturally, the reverberation radius is strongly correlated with the reverberation time. We also computed the reverberation radius, so the sound samples could be labeled as either clean or reverberant. We regarded clean speech as speech measured inside the reverberation radius and reverberant speech as speech outside the reverberation radius.

The parameter values used in the Eyring-Norris equation were used as input to the shoebox model, which simulates an impulse response in a rectangular room, a shoebox. The shoebox model is an implementation of the image source method of Allen and Berkley (1979). The speaker and the listener or microphone are modeled as two points in space. Apart from the direct sound, specular reflections are computed using mirrored image sources. An impulse response is obtained for every image source. The final impulse response describing the room is computed by combining all individual impulse responses, which are received at different delay times. This impulse response is convolved with the speech signal, resulting in reverberant speech. The speech is processed with a gammachirp filter bank (Irino and Patterson, 1997), which results in a logarithmic time-frequency representation (a cochleogram, see appendix B). Figure 2.1 depicts a cochleogram of a clean speech sample on the left, and a cochleogram of a reverberant speech sample on the right.

We expected that the effect of reverberation on speech can be measured directly in the audio signal. Since we want to test whether we can measure reverberation in an uncontrolled environment, the features used for the classification must have no parameters that require knowledge of the room characteristics. One prominent effect of reverberation on speech is the attenuated salience, that is, the attenuated stability in both the frequency and the energy, of the harmonics (Darwin and Hukin, 2000). Therefore, voiced speech was located based on the selection of harmonic

Table 2.1: Features that indicate the reverberation level on a harmonic track (h). $E_h(t)$ is the energy development and $f_h(t)$ is the frequency development of a harmonic track in time. MA is the moving average of an energy or frequency track and P is its polynomial. See appendix C for the calculation of the features.

Energy variation		Harmonic energy salience		Harmonic frequency salience	
peak rate (PR)	(2.2a)	$\Delta E_h(f)$	(2.2c)	Var f_h/MA	(2.2f)
Var E_h	(2.2b)	Var Δf_h	(2.2d)	Var f_h/P	(2.2g)
		Mean Δf_h	(2.2e)		

complexes—a superposition of co-occurring harmonics—in the cochleogram, using the algorithm presented in Krijnders *et al.* (2007). Subsequently, the fluctuation in energy and frequency of the first five harmonics of the harmonic complex was measured. These harmonics can be better resolved from the cochleogram because of the logarithmic frequency scale, and are hence more reliable.

We measured the energy and frequency fluctuation through seven features, which are summarized in Table 2.1. The energy variation was measured through the peak rate of the harmonic track (2.2a) and the energy variation of the harmonic compared to its smoothed version (2.2b). Both values are expected to increase at higher reverberation levels. In addition, the energy contributions of echoes cause less distinct harmonics. This effect was captured by calculating the energy slope of the harmonics (2.2c), and the variation (2.2d) and mean (2.2e) of the width of the harmonic compared to an ideal sinusoid. In other words, the energy slope is less steep in reverberant speech, and the harmonic covers a broader frequency range. Reverberation effects can be found in the time-frequency space as well. The short-time development of the harmonic is distorted by echoes, causing a less smooth harmonic track. Therefore, the track variation was measured compared to its smoothed version (2.2f), and to an approximation of a clean harmonic track (2.2g). The calculation of all seven features is worked out in appendix C.

2.4.2 Experiment

Part of the Aurora database (Hirsch and Pearce, 2000) was used to validate the six features. Artificial reverberation was added to 685 randomly selected clean sound

samples with a mean duration of 1.5 seconds, spoken by 214 different speakers, both male and female. The reverberation was computed at nine levels, equivalent to reverberation times distributed roughly linearly between 0 and 1600 milliseconds. As we expected, most of the 685 sound samples showed a significant correlation of at least one feature with the reverberation time. Only 6% did not show a relation for any of the features. However, the predictive strength of the features for individual sound samples is no direct indication for the general classification of the speech samples as either clean or reverberant. To test classification, global thresholds need to be determined in a training set and used to classify a test set.

Since the 685 speech samples were reverberated at nine levels, a total of 6165 samples could be used for classification. After the dismissal of samples in which we could not measure one or more features[1], 5189 samples were left. All samples within the reverberation radius ($T_{60} = 0.22$ seconds), the two lowest levels, were labeled as clean, and all samples outside the reverberation radius, the other seven levels, were labeled as reverberant. The data was split in a part for training (33%) and a part for testing (66%). In addition, continuous read speech of six speakers was recorded using a close-talking microphone. This data was split into samples of similar length to the Aurora database, and resampled to an equal sample frequency. The speech samples were artificially reverberated in the same way as the other data. Again, part of the data that was unfit was removed, and 2377 samples with a mean duration of 2 seconds were left. These samples served as an extra test set, which can show the robustness of the features. Finally, both data sets were also labeled with a different threshold for the reverberation time ($T_{60} = 0.7$ seconds) to test the performance on a more balanced design.

Numerous methods exist to test the classification accuracy of features. We used a support vector machine (SVM), since it is known to be less prone to problems of overfitting than some other methods (Duda *et al.*, 2001). In training, an optimal separating hyperplane, or threshold boundary, is determined. The support vectors are the speech samples that are closest to the hyperplane, and hence are most difficult to classify. The mapping of the data to a higher-dimensional space is dependent on the type of kernel, that is, the mapping function, which can be defined by the

[1] In these samples the harmonic complexes were not sufficiently salient to be segregated by the algorithm.

user, or selected from one of the standards. For our data we use a standard linear kernel. The number of support vectors is an indication of the complexity of the classification. During the testing phase, the speech samples are mapped onto these support vectors. Since the test samples are labeled as well, the classification can be compared to the labels, resulting in a performance measure.

2.4.3 Results

The speech samples from the Aurora database were split randomly into a training set of 1744 samples and a test set of 3445 samples. The seven features (one of which, $\Delta E_b(f)$, has two values) were computed on the first five harmonics, resulting in 40-dimensional data. The skewedness of the first data set—22% of the speech samples was clean and 78% was reverberant, because the reverberation radius corresponds to a relatively low reverberation time of just over 200 milliseconds—is accounted for by using prior probabilities to weight the class error contributions. The SVM was trained on the training samples, resulting in a classifier with 363 support vectors. The rest of the speech samples of the Aurora database was tested on the trained classifier. The performance, or accuracy, of the classifier was 92%. The additional speech samples of our own recordings were also tested on the classifier, with a performance of 87%, only a few percent less. The same procedure was applied to the second, balanced, data set, resulting in an accuracy of 80% on the Aurora samples, and 70% on the recorded speech samples. The results are summarized in Table 2.2. Figure 2.2 gives the detailed results of one of the results (the Aurora data with a labeling threshold of $T_{60} = 0.7$ seconds). The classification of the samples is skewed toward the more reverberant samples, because the features do not develop linearly with the reverberation time. In a separate experiment we tested the predictive strength of the features, and found that they overestimate the reverberation time below approximately $T_{60} = 0.8$ seconds, while the differences between higher reverberation times cannot be predicted (see appendix C). This effect is explained by the difference in the effect of reverberation on the audio signal, which is greater for smaller reverberation levels.

Different classification methods could be chosen for this problem, or different settings for the SVM. For example, if the size of the training set is increased, the performance on the other Aurora speech samples increases, but the performance

Table 2.2: Results of the SVM classifier on two data sets (Aurora and continuous speech) with two designs (unbalanced for the reverberation radius R_{reverb} and balanced).

T_{60} threshold	Accuracy Aurora data	Accuracy recorded data
0.22 sec. (R_{reverb})	92%	87%
0.7 sec.	80%	70%

on the extra test set decreases. We are not interested in optimizing the classifier on a particular data set, but in the separability of any reverberant speech using the features. Hence, the classification with an SVM using a linear kernel gives an indicative performance result.

2.4.4 Discussion

Although many methods to assess or resolve the effects of reverberation are successful in improving ASR, they generally cannot be used for different applications. Since these methods are designed for ASR, they utilize the common assumptions of ASR that cannot necessarily be met in an uncontrolled environment. First, methods that adjust the training conditions to reverberant conditions or apply inverse filtering cannot deal with variable or unknown conditions. For example, the model of Couvreur and Couvreur (2004) is not tested on speech signals that are affected by transmission effects other than reverberation, such as background noise or concurrent sources. Second, blind dereverberation methods based on spectral subtraction and blind reverberation classification methods like those presented in this section rely on the presence of speech to estimate the reverberation components (Wu and Wang, 2006; Kinoshita *et al.*, 2009). However, if a system must recognize sound events in a continuous audio signal, no assumptions about the presence of a specific sound event can be made. Therefore, dereverberation methods should be extended to include estimation of reverberation for impact sounds and broadband sounds, not only for tonal sounds.

In general, if researchers want to improve sound event recognition in unknown conditions, they should focus on developing robust techniques for online blind dereverberation of mono-signal input with unrestricted content. More specifically, the common experimental paradigm for resolving distortions should extend to

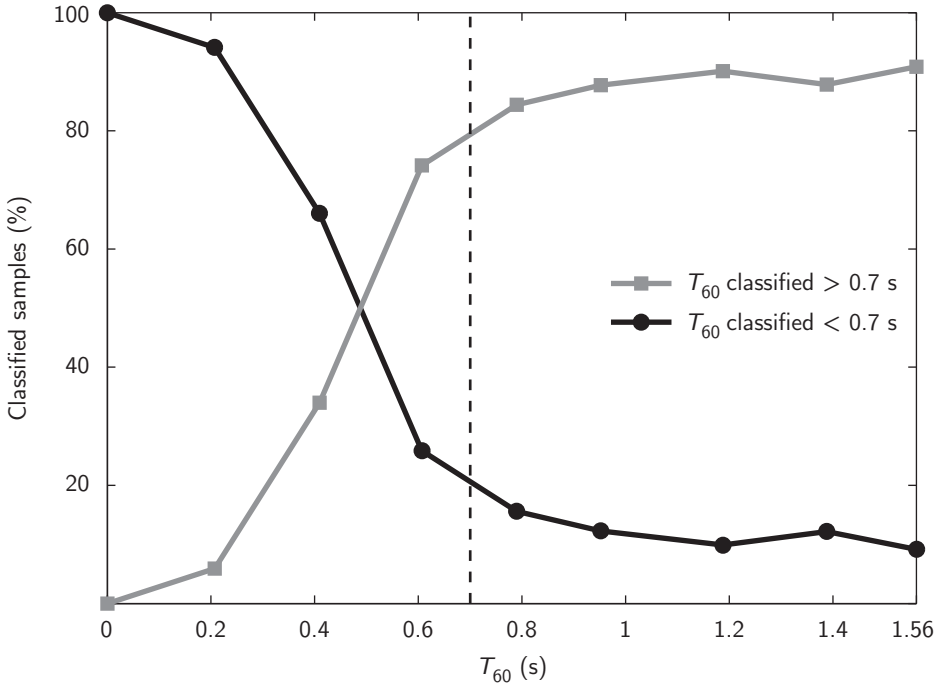


Figure 2.2: Detailed results of the SVM classifier on the Aurora data set with a labeling threshold of $T_{60} = 0.7$ seconds (indicated by the dashed line). The overall accuracy on this data is 80%. The lines indicate the percentage of sound samples that were classified as either below the threshold ($T_{60} < 0.7$) or above the threshold ($T_{60} > 0.7$) for each reverberation level.

conditions outside of ASR. Methods to assess or resolve the effects of reverberation should be tested in real-world environments, that is, with continuous audio recordings in unknown and possibly variable conditions. When the experimental paradigm is shifted toward these challenging conditions, the development of techniques for audio quality improvement needs to focus on robustness instead of perfection in limited domains.

2.5 CONCLUSION

In this chapter we have explained that automatic sound recognition in real-world environments requires sound event segregation that works in variable environ-

ments, that is, environments with different sorts and levels of transmission effects and varying co-occurring sounds. Furthermore, we have discussed that current techniques for automatic sound recognition are mostly designed for specific applications, such as speech recognition, and not for general purpose sound event recognition in unconstrained environments. In other words, the semantics of the sound are given by the application. Therefore, they can rely on signal-driven methods (based on the acoustic properties of the signal) to deal with transmission effects within a single domain. However, in real-world environments the semantics or context of sound events are not stable or even known. Instead of assuming a specific context, we argue that the context can be learned, and used to manage unreliable signal information. In fact, robust general purpose sound event recognition is infeasible with signal processing techniques only, because they can at most provide hypotheses of components that are likely to constitute a single sound event. Selecting and identifying the target sound event is only possible by means of non-acoustic factors, such as the goal of a system.

Moreover, even if researchers can assume or have accomplished perfect sound event segregation and classification for real-world sounds in a similar way as in speech and music processing, the meaning of the classified event is not yet known. When people listen in every-day life, they give meaning to the events that they hear. This meaning is not only based on acoustic properties or class membership. People use their memory, experience, and expectancy to give meaning to their (sonic) environment. Hence, these factors influence what is heard (segregated). In the next chapter we discuss the formation of the human percept of a sound event, and the role of context in this process.

3

Context in human auditory cognition

The content of section 3.3 has been published as part of Niessen, M. E., Van Maanen, L., & Andringa, T. C. (2008). Disambiguating sound through context. International Journal on Semantic Computing 2(3), 327–341. Proceedings of the 2nd IEEE International Conference on Semantic Computing, 88–95.

In the first part of this chapter we give background on the perception of sound events, which is often referred to as auditory object formation, in analogy with visual object formation. However, instead of auditory object we use the term auditory event, because it captures the dynamics of sound. Auditory events can be described by a set of properties that are related to either constancy or separability of the auditory event. Different approaches have been taken to study auditory events and to explain their properties. However, researchers should be aware of the exact process they are studying. Mostly, the stimuli used in perception studies are artificial and simple, so that the perceptual process of streaming and segregation can be studied in a controlled setting. Although the earlier stages in the auditory process are important in auditory event formation, they do not constitute the complete phenomenon. Cognitive processes, such as attention and memory, play a dominant role as well. In the second part of the chapter we present an experiment to show the effect of context in sound event recognition. The results of this experiment indicate that context can facilitate sound event recognition. We do not attempt to explain the cognitive processes that underly the influence of context on auditory event formation. Instead, we aim to demonstrate the importance of other factors beyond basic acoustic properties of the sound in auditory event formation.

3.1 INTRODUCTION

In the previous chapter we have discussed the challenges of an automatic sound recognition system in real-world environments. State of the art systems mostly restrict the search space, either of the input or of the operating environment, to function reliably. In contrast, people have no difficulties in recognizing sound events in many different and noisy situations. For example, they can have a conversation surrounded by other people talking, a phenomenon called the cocktail party effect (Cherry, 1953; Bronkhorst, 2000). This ability relies on the bidirectionality of human perception. People use their experience, attention, and knowledge of the world to give meaning to a sound (Box 3.1) as well as signal-driven (bottom-up) strategies. Because our aim is to automatically recognize sound events, we discuss how humans form the percept of a sound event (the auditory event) in section 3.2, based on a survey of studies about auditory event formation.

One important factor that allows people to hear in an unconstrained environment is their knowledge of the context, which helps them to form predictions and guide their perception of the environment (Bar, 2007). Events in the real world generally do not occur in isolation, but co-occur with other events and particular environments (Oliva and Torralba, 2007). Therefore, the meaning or the semantics of a sound event is influenced by the associations that people have to other events and environments. In the following, whenever we talk about *context*, we refer to it as the learned associations of an event to environments and co-occurring events (Box 3.2, page 35). In section 3.3 we present an experiment to test how context facilitates sound recognition. Finally, in section 3.4, we substantiate how an understanding of human sound event recognition can help in automatic sound event recognition.

3.2 AUDITORY EVENTS

When people are asked to describe a sonic environment they will describe the different sound events in terms of the sources that caused the events (Ballas, 1993; Vanderveer, 1979). They will normally not describe the acoustic properties of the sound events. For instance, a passing car will be referred to as a car, not as a noisy harmonic complex in combination with a burst of noise. The evaluation of sound events in terms of the sources or processes that produced them is often named

Box 3.1: Meaning

For systems one cannot talk about the meaning of something in the same manner as we can for people, which is demonstrated by Searle's famous Chinese room argument (Searle, 1980). Briefly, his argument consists of imagining a person doing the Turing test (Turing, 1950) in Chinese. The person receives input symbols, performs manipulations based on their shape and some provided rules, and returns output symbols. Suppose these output symbols are indistinguishable from what a real Chinese speaker would reply. Does the person understand Chinese? Obviously not, because he cannot read Chinese, and does not interpret the symbols. The symbol grounding problem (Harnad, 1990) defines this as the problem that the meaning of a symbol is not intrinsic to the symbol, but is given its meaning by a person. In other words, an algorithm that manipulates symbols systematically to generate an output based on their properties does not give meaning to the symbols. Therefore, one cannot directly compare the output of a sound recognition system to the interpretation of humans. However, efforts have been made to substantiate a *semantic* interpretation or representation independent from a human mind (Fodor and Pylyshyn, 1988; Newell, 1982). A system can detect instances that correspond to some event or object in an input signal, categorize and identify them based on their properties, and act according to a learned semantic interpretation. For example, a system can learn semantically related concepts based on their statistical relationship in training data (latent semantic analysis, Landauer and Dumais, 1997). Furthermore, past experiences can lead an automatic agent to act according to a maximized utility (Kaelbling *et al.*, 1996) or affect (Schermerhorn and Scheutz, 2009). In conclusion, the meaning of a sound event is not inherent to the event, but given by a human listener. Although a system cannot give meaning to a sound event, models or algorithms can be designed to use similar strategies as humans and reply in a consistent way, comparable to humans.

everyday listening (Gaver, 1993). Everyday listening relies on the ability of the human perceptual system to segregate parts of an auditory stream into different elements that might represent individual events. Yost (1991) distinguishes fusion and segregation as the two components that constitute the potential of event recognition. Fusion refers to the grouping of sound components of an event into a single representation, the auditory image. Segregation, or auditory streaming, refers to the separation of different auditory images from each other. Finally, the auditory images are classified as particular sound sources. This way of relating to the process of event recognition can be connected with the way Shinn-Cunningham (2008) refers to it as (bottom-up) auditory object formation. However, she also stresses the importance of top-down attention on object formation, while Yost (1991) considers auditory image formation as a unidirectional process.

The term auditory object is usually preferred over auditory image, because an image is associated with the visual representation of a sound, for example a time-frequency representation (Shamma, 2001). Furthermore, the term object carries a sense of wholeness. However, it can bias ones interpretation toward a static thing instead of an event, because it stems from visual research.[1] Furthermore, there has been some debate about how an auditory object is defined (Griffiths and Warren, 2004). For example, in cognitive neurophysiological experiments auditory objects are mostly equated with artificial units, like (combinations of) tones (Atienza *et al.*, 2003; Dyson and Alain, 2004; Winkler *et al.*, 2006). In the field of environmental sound event recognition these artificial stimuli are of less interest, because they do not occur in real-world environments. Instead, the concept we choose for the human percept of a sound event should relate to the dynamic events that occur in the everyday environment of a listener. Therefore, we refer to it as an *auditory event*. Similarly, we use *auditory episode* instead of auditory scene to refer to the mental representation of a sonic environment. Figure 3.1 shows a schematic overview of the relation between some important concepts we use throughout the chapter: sound source, *physical action*, sound event, sonic environment, auditory episode, and auditory event.

[1] Dictionary entries for "object" are variations of "a material thing that can be seen and touched".

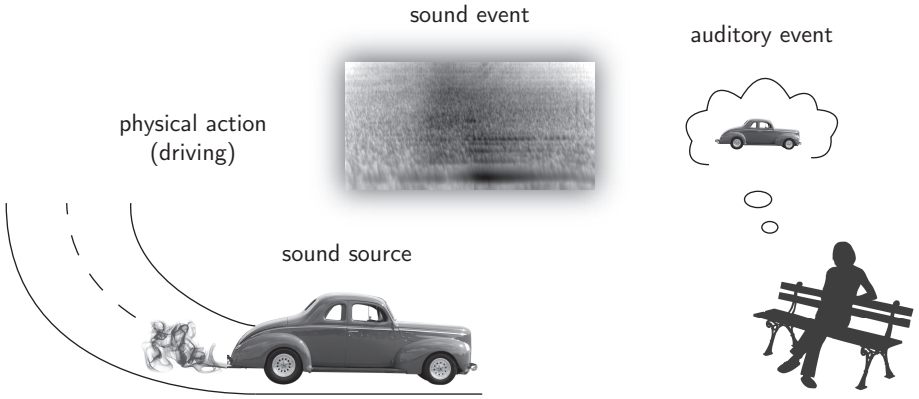


Figure 3.1: Example to demonstrate the relation between some important concepts. A possible sound source, such as a car, that is involved in a physical action, such as driving, results in a sound event. If this sound event is perceived by a human listener, his cognitive representation of this event is called an auditory event. We refer to a collection of different sound events as a sonic environment, and a cognitive representation of the sonic environment is an auditory episode.

3.2.1 United approach to auditory events

Different areas of research use different methodologies to study a phenomenon, such as auditory event formation. However, researchers should be aware that the meaning of a concept is different when it is described in different domains, and is studied with different methodologies. A risk of degradation of the concept of a studied phenomenon arises if it is not defined, especially when a concept that originates from one field is transfered to another field. For example, since the methodological progress in neuroscience, many studies have been conducted that map psychological attributes to areas in the brain. Regularly, in these studies brain areas are said to see, feel, and so forth. However, it does not make sense to ascribe a psychological predicate about a person (a whole) to his brain (a part) (the mereological fallacy, Bennett and Hacker, 2001). Although auditory event formation is not a psychological predicate, it is a perceptual predicate, which describes a trait of a person as a whole.

To avoid the mereological fallacy, researchers that model or describe auditory or brain processes should be aware that they do not describe or explain the au-

ditory event as it is perceived by a person. Instead, they study and describe the processes that are involved in auditory event formation at different abstraction levels. Dennett (1991) discusses how folk-psychology and the patterns it predicts—Dennett (1987) defines a pattern as some part of behavior that is predictable assuming intentionality—relate to the physical level. By means of Conway's Game of Life Dennett step by step illustrates how at different levels of description there are different sets of patterns by which one can make predictions. At lower levels it is costly and hence difficult to make predictions, but the predictions are correct. As the level of description gets more abstract, it becomes easier to make predictions, but there is also more noise, so more mistakes. "Predicting that someone will duck if you throw a brick at him is easy from the folk-psychological stance; it is and will always be intractable if you have to trace the photons from brick to eyeball, the neurotransmitters from optic nerve to motor nerve, and so forth" (Dennett, 1991, p. 42). Analogously, auditory event formation can be studied at different levels of description.

The different description levels of auditory event formation range from the physical description of a sound event to cognitive models that explain the role of attention. Griffiths and Warren (2004) propose a similar approach, in which complementary models produce testable hypotheses that explain auditory event analysis. They focus their framework on auditory pathways and regions in the cortex studied within psychophysics and neuroscience. The studies within these fields model small parts of the process of auditory event formation. Therefore, they can be said to describe it at lower description levels, at which the predictions are difficult to make, but precise. However, perceiving an auditory event is an experience of a person. Therefore, methodologies from fields such as cognitive psychology can enhance the understanding of auditory event formation as well, at higher description levels.

3.2.2 Constituents of auditory events

Auditory event formation can be studied at many different description levels, because it is influenced by many physical, perceptual, and cognitive factors. To understand the whole process, we should structure and define the concept of an auditory event and its constituents. Bregman (1990) makes a distinction between

primitive auditory scene analysis (ASA) and schema-based ASA to structure hearing. Primitive ASA refers to the signal-driven analysis of properties of the sound, such as grouping and segregation, while schema-based ASA accounts for the cognitive schemas that influence the perception of the grouped and segregated auditory event. However, the boundary between these two types ASA is arbitrary, or even superfluous, because the effect of cognitive factors can influence the primitive processes, and vice versa. For example, several studies in neuroscience have shown that attention can influence auditory streaming (Cusack, 2005; Carlyon, 2004). Furthermore, brain imaging studies have demonstrated that auditory cortical areas are active during primitive processes, which imply a tight coupling between primitive and cognitive processes (Gutschalk *et al.*, 2005). Because the two types of scene analysis are intertwined, we adopt a different framework to analyze the process of auditory event formation.

While some research areas limit the concept of an auditory event to perceptual grouping principles (for example Jones *et al.*, 1998), our concept encompasses all the stages of processing, from the perceptual analysis of the sound, to the cognitive processing. Griffiths and Warren (2004) give four principles of event analysis that guide our interpretation of auditory event formation, summarized in Table 3.1. First, an event has a relation to sensory information, so an auditory event is based on hearing something in the world. Second, the event is separated from the rest of the sensory information. In other words, an auditory event is segregated from other auditory information. Third, the perception of the event is invariant over different experiences. Hence, the distinguishing features of an auditory event can be generalized over different conditions, like reverberation or background sounds. Fourth, an event is not necessarily bound to one sense. For example, a car is not perceived as a different event when it is either seen or heard.

To comply with the first principle we should further clarify what an auditory event is (Figure 3.1). Hearing something in the world implies that the representation, the auditory event, is based on a process that is produced by a physical source (Gaver, 1993). Whether the auditory event refers to the source or to the process depends on how it is categorized or described (Dubois, 2000; Griffiths and Warren, 2004; Guastavino, 2007). Information related to the source, such as a car, is specified by the properties of the source. On the other hand, information related to the process, such as accelerating, is specified by the patterns of change. Depend-

Table 3.1: Auditory event (Griffiths and Warren, 2004)

Properties of an auditory event:
1. Based on information from sound events
2. Separated from information about the rest of the sonic environment
3. Generalized over different experiences
4. Generalized over different senses

ing on the goal of a person the auditory event can refer to either the source or the process or a combination of both. Furthermore, the type of source or process also influences what type of information the person focusses on. In an experiment on free environmental sound naming Marcell *et al.* (2000) found that listeners name a sound by the source object or action when the agent is human ("bowling", "harmonica"), while animal sounds were consistently referred to by the agent ("cat").

The second principle (separability) has a long history in perceptual research, and many different viewpoints. Van Valkenburg and Kubovy (2004) differentiate three approaches: auditory events, auditory streams, and figure-ground segregation. The first approach focusses on the distinction between sounds and events (Rosenblum, 2004; Blauert, 2001). Events in the world structure the sound. Therefore, people do not hear properties of the sound, such as pitch, but events, such as driving cars or singing birds (Gaver, 1993). The second approach is based on the Gestalt principles of visual perception (Köhler, 1967), and described for sound by Bregman (1990). Elements in the audio signal are combined by the perceptual system into streams, based on principles such as proximity and continuity (an example is shown in Figure 3.2). The third approach is an analogy of the result of figure-ground segregation in vision (Figure 3.3). Perceptual units of attention, corresponding to objects, are separated from the background (Carrell and Opie, 1992; Scholl, 2001). The formation of these objects is guided by perceptual organizations, or gestalts, of the audio signal, which are the presumed objects. People attend to one (or several) of these objects, while the ground remains undifferentiated (Kubovy and Van Valkenburg, 2001).

The third (and fourth) principle (constancy) refers to the ability of human listeners to retrieve abstract information about a sound event that is independent of

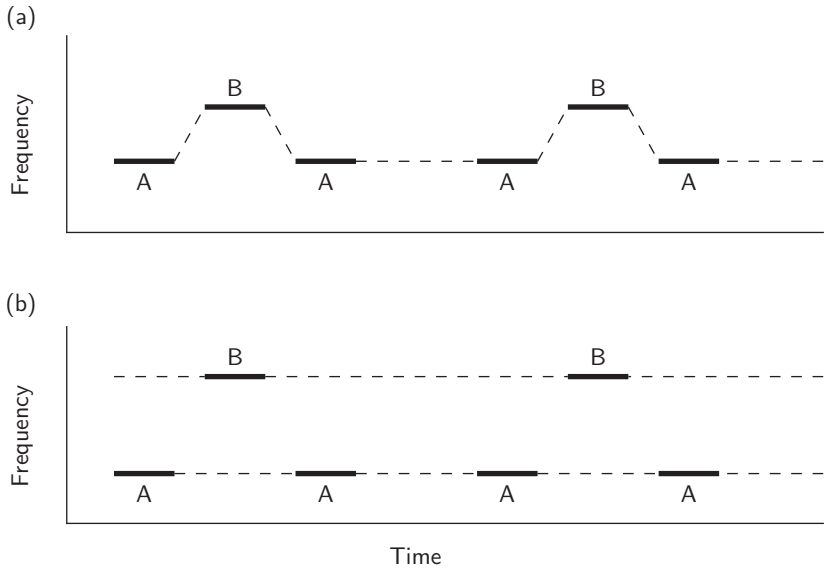


Figure 3.2: In this tone sequence a person can hear either one stream with the alternating tones ABA ABA (a) or two tone streams A A and B B with different rithms (b), depending on the proximity in frequency of the two tones (Van Noorden, 1975).

the modality. The invariant perception of physical properties of sound events is the focus of the ecological approach to perception (Gibson, 1966). Many studies in the field of ecological hearing have been aimed at showing the perception of a specific physical property of objects in different situations (Warren and Verbrugge, 1984; Carello *et al.*, 1998; Kunkler-Peck and Turvey, 2000). Their key point is that the (physical) information about a process or object is present in the signal it transmits. A different (although not incompatible) approach is the information processing approach from cognitive psychology, which analyzes cognition and perception by abstract stages in the processing of a task (Anderson, 2005). Auditory perception can be analyzed as a succession of conceptual processing stages, from sensory transduction, via auditory grouping and categorization, to recognition (McAdams, 1993). According to this approach, which is more functional than the ecological approach, the constant perception of events over different experiences can be explained by theories about memory. People structure the world in categories (Rosch, 1975; Dubois, 2000), and memorize a prototype (Reed, 1972; Smith and Minda, 2000)

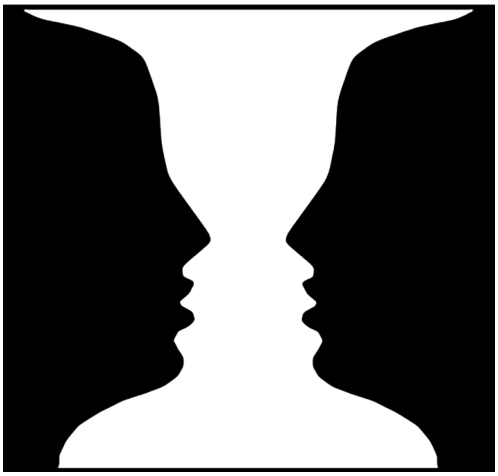


Figure 3.3: Human perception separates objects (figures) from their background (ground), based on their interpretation, which is supported by properties of the objects, such as borders and depth. In Rubin's vase (or face) the image can be perceived as either two faces or as a vase. Because the figures share their borders, they cannot be the figure at the same time.

or exemplars (Medin and Schaffer, 1978; Nosofsky and Zaki, 2002) of a category.[1] Whenever they perceive a new instance of an event or object they try to match it to a prototype or exemplar in memory (Figure 1.4).

As we indicated in the introduction of this chapter, we favor the term event over object, because it captures the dynamics of sound events. The term auditory event is chosen by the ecological approach to indicate direct perception of physical actions (Fowler, 1996), but our concept is broader. An auditory event is a representation of a sound event in memory. Therefore, it is not necessarily initiated by a physical action. In other words, an auditory event can be a prototype instead of an exemplar of a sound event.

[1] Prototype models and exemplar models are two major theories of category representation. Prototype models assume a category is represented by an abstract prototype, while in exemplar models categories are represented by (good) exemplars of that category.

Box 3.2: Context

The dictionary entry (Merriam-Webster) for context is "interrelated conditions in which something exists or occurs". The term is often used in perception and memory literature to refer to non-target information, for example a visual scene surrounding some target object in an identification experiment. Because the term context is so widely used, we will give a short survey of its application in different research domains. In this setting we can clarify our use. In memory research context is usually equated with the associations that are triggered by perceiving something in the world (Bar, 2007; Oliva and Torralba, 2007). These associations are learned through experience. They can refer to events, objects or environments, that is, they are not bound to one sense or one type of thing. In contrast, in visual perception research, context refers to the visual scene in which a target object is presented (Palmer, 1975; Hollingworth, 1998). If the target object semantically fits the scene, the context is called consistent or appropriate. If the target object does not fit the scene, the context is inconsistent or inappropriate. In experiments in which the effect of an (in-) consistent visual scene on object recognition is tested, both the object and the visual scene are usually drawings. In speech perception context is the linguistic information (Ganong, 1980). For example, if a speech sound is impoverished, the sentence of which it is part helps to recognize the speech sound. In auditory perception research aimed at sounds other than speech, context is used less consistently. One important reason is that a sonic environment, which is dynamic, is more difficult to represent than a visual scene, which can be represented statically. For example, Ballas and Mullins (1991) presented context as a sequence of sound events, while Gygi and Shafiro (2006) mixed the target sound event with recorded sonic environments. All applications of context in visual and auditory perception experiments have in common that a (schematic) representation of a surrounding is presented to participants in the same modality as the target. Therefore, the context as it is presented in experiments is more limited than in a real-world environment. We define context as the learned associations of an event to other events and environments (as it is defined in research about human memory). However, in an experimental setting we may be restricted to certain aspects of context. In this case we will indicate which aspects of context we are using.

3.3 HUMAN SOUND EVENT RECOGNITION

Since we want to model the role of context (Box 3.2) in sound recognition, we are also interested in its role in human sound recognition, which has received little scientific attention (Gygi and Shafiro, 2007).[1] Therefore, we present the results of an experiment that has been designed to determine whether context facilitates the interpretation of an ambiguous sound event. It is known that sound events are more difficult to recognize when they may stem from multiple types of sources (Ballas and Howard, 1987). Context is essential to disambiguate these sound events, as is shown in an example of the same study. In this example, participants interpreted a sound event differently when it was combined with another sound event and different instructions. A follow-up study did not find this facilitatory effect, but did find a suppressive effect of an incongruent context (Ballas and Mullins, 1991). Similar results, that is, both a facilitatory effect of context (Palmer, 1975), but also the lack of it (Hollingworth, 1998) have been found in visual object recognition, although the general consensus is that the context does help to recognize objects (Bar, 2004). These results show that context is a complex factor. Moreover, context can be perceived in many different ways, such as in sound and image, but also in time of day and place of occurrence.

The experiment described in this section is designed to show one particular effect, namely the facilitatory effect, that context can have on the interpretation of an ambiguous sound event. The results of the experiment will be important for automatic sound recognition in two ways. First, if context is shown to be beneficial for human recognition of ambiguous sounds, it can also be useful in an automatic system that needs to recognize ambiguous events. Second, applications of real-world sound event recognition, especially those that need to interact with a user, can benefit from having a representation of the environment that is comparable to a human listener.

To test the facilitatory effect of context in human sound event recognition we presented homonymous sound events to participants. Homonymous sound events are defined by having two (or more) possible interpretations, like one word can refer to multiple concepts. When these sound events are presented in isolation,

[1] Vision has been and is the focus of perception research. For example, if one searches for articles with "vision", Google Scholar returns about twice as many results as for the query "hearing".

the probability that they are identified as any of their possible interpretations is equal. In contrast, when homonymous sound events are preceded by a sound event that predisposes the listener to one of the two interpretations, we expect a biased response toward that interpretation. For example, the sound of a purring cat can be ambiguous without any context information, because some engines make a similar sound. However, if a person is first presented with a context sound event, such as honking, it is more likely that he will recognize the sound event as an engine than as a purring cat. Hence, in this experiment context is defined as the sound event that creates a sequence of events, instead of an isolated sound event. In other words, the context sound event can trigger associations for the participant that influence the recognition of the target sound event.

3.3.1 Method

To create homonymous sound events we used pairs of similar sounds from high-quality commercial sound effects recordings (Hollywood Edge and Sound FX The General), which were used previously to study the similarity of sound events (Gygi *et al.*, 2007). Sound pairs that were found maximally similar in this study were combined to form chimaeric sounds. Chimaeric sounds are composed of the fine time structure of one sound and the temporal envelope of another sound (Smith *et al.*, 2002). The signal properties of the sound events varied greatly because of the diversity of the environmental sounds in the database. Hence, the chimaeric sounds did not always result in homonymous sound events. For 12 selected homonymous pairs, listed in the left part of Table 3.2, we chose the combination of fine structure and envelope that sounded most natural to the experimenter. Most of the envelopes of sounds A were used for the chimaeric sounds, while most of the fine structures of sounds B were used. The homonymous sound events had a mean duration of 2.8 seconds. The sounds that provided context for the homonymous sound events, listed in the right part of Table 3.2, were obtained from additional commercial recordings (Auvidis and Dureco). All sounds were sampled at 44.1 kHz. The total of 52 sound event sequences (two context conditions for the homonymous sound events, and 28 filler sequences, see next paragraph) had a mean duration of 7.7 seconds. The context sound events preceded the target homonymous sound events such that the sequence sounded most natural. However, the context sound event

Table 3.2: List of similar sound pairs used to form homonymous sound events (left) and the context sound event that facilitated them (right).

Sound event		Context	
A	B	A	B
Pouring water	Rain	Refrigerator door	Thunder
Thunder	Passing airplane	Rain	Airport PA
Whistle	Singing bird	Football cheering	Forest
Footstep	Drum	Closing door	Guitar
Toilet flush	Pouring water	Urinating	Refrigerator door
Meowing cat	Crying baby	Barking dog	Music box
Coughing	Barking dog	Talking	Meowing cat
Basketball	Closing door	Cheering crowd	Footsteps
Ticking clock	Pingpong ball	Chiming clock	Applause
Water bubbles	Horse running	Teakettle whistle	Horse neighing
Bowling	Thunder	People talking	Rain
Zipper	Car starting	Raining on tent	Car door closing

always ended before the end of the target sound event. For example, when the context sound event was rain, it continued through the start of the sound of thunder, but when the context sound event was the closing of a refrigerator door, it ended before the sound of the pouring water started.

In total 42 participants with a mean age of 24 took part in the experiment. Six participants reported a slight hearing loss, but showed no decrease in their performance on the filler sounds compared to the normal hearing participants.

The experiment comprised three conditions, one in which the context sound event facilitated the interpretation of sound event A, one in which the context event facilitated the interpretation of event B, and a control condition in which the target sound events were heard in isolation. The three conditions were presented between the participants. The homonymous target sound events were alternated with 28 filler sound events taken from the same database. They were included to assess the performance of the participants, and to make the participants unaware of which sound events were the targets. The total of 40 sound events was presented in random order, but no targets were present in the first 6 exposures to familiarize

the participants with the task. The recognition task was a binary choice task. For the target sound events the participants could choose between the descriptions of the two original sound events, and for the filler sound events they could choose between the actual cause and some other related source description. Furthermore, the participants had to indicate on a four-point scale how confident they were of their answer. The control group of 11 participants heard the sound events in isolation. The second group of 15 participants first heard a sound semantically consistent with context A followed by the target chimaeric sound. Finally, the third group of 16 participants first heard a sound semantically consistent with context B followed by the chimaeric sound. The 28 filler sequences, the filler sound events preceded by a semantically consistent sound event, were the same for the last two groups. The control group heard the filler sound events without a context event. The experiment was conducted online during January 2008.

3.3.2 Results

The score of all participants in every group on the filler sound events was 100%, and they gave a mean confidence rating of 2.8 on a four-point scale ranging from 0 to 3. A two-way analysis of variance (ANOVA) was used to test the difference in the response between the participants within the homonymous sound events. The effect of context A on the mean recognition score compared to the mean score in isolation was significant: $F_1(1,11) = 8.09$, with $p < 0.017$. However, there was no effect of context B on the mean recognition score compared to the mean score in isolation ($F_1(1,11) < 1$). The results are summarized in panel (a) of Figure 3.4. The black bars depict the average score on option A for all participants within a group summarized for all homonymous sound events, where option A is the event description that is in agreement with context A. The complement, 100% minus score A, is the average score on option B (the gray bars).

The difference between the confidence ratings in correct responses, that is, responses for which the answer was in agreement with the context sound event, compared to the confidence ratings in incorrect responses was significant in the group that heard context A: $t(101) = 3.34$, with $p < 0.002$. The confidence rating was higher when the answer was in agreement with the context. The mean confidence ratings of consistent and inconsistent recognitions are depicted in Figure 3.4 (b).

This effect was absent in the group that heard context B ($t(159) < 1$).

Not all chimaeric sounds appeared to be as homonymous as assumed. In particular three sound events received one interpretation exclusively in the isolated condition. When these three sounds were excluded from the ANOVA, the difference in the mean score of context A compared to the mean score in isolation had a greater F: $F_1(1,8) = 13.28$, with $p < 0.007$. In conclusion, for the homonymous sound events we found a significant effect of one context on recognition.

3.3.3 Discussion

Although there is a significant effect of one context on the mean scores, this effect is completely absent in the other context. The explanation for the absence of the effect lies in the design of the experiment. The homonymous sound events were formed by combining the envelope of one sound and the fine structure of another sound. Most descriptions of context A predisposed the participants to the interpretation of the envelope of the homonymous sound, while the interpretation related to the fine structure was most prominent in context B. Hence, the envelope is a stronger cue for recognition than the fine structure for this experimental design. This effect is known in speech perception (Shannon *et al.*, 1995; Smith *et al.*, 2002), and depends on the number of frequency bands used to create the chimaeric sound. If the number of frequency bands we used (eight) were used for the recognition of chimaeric speech sounds, the fine structure would give relatively little information compared to the envelope. Hence, our results suggest this effect can be generalized to environmental sounds. As a consequence, the effect of context is canceled by the preference for the envelope in context B. This conclusion is consistent with a significant prevalence for the interpretation that coincided with the envelope of the homonymous sound event (64%) compared to the fine structure (36%) when the sounds were presented in isolation ($\chi^2(1) = 9.82, p < 0.002$). Overall, the experiment demonstrates that the context in which a sound event is heard constraints its perception.

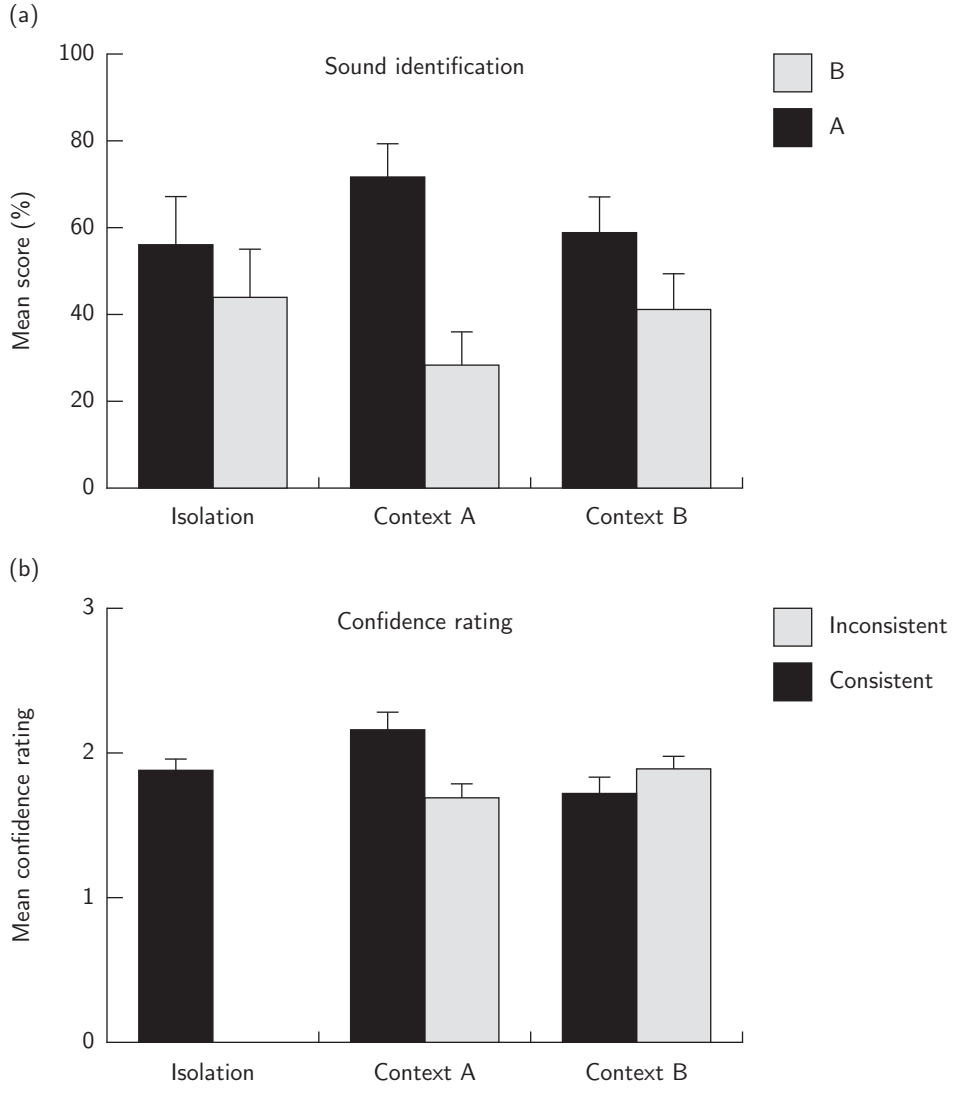


Figure 3.4: Panel (a) shows the mean scores on the option consistent with both context A and B in each of the three groups, with the standard error. The sum of both scores in each group add up to 100%. The mean confidence ratings of consistent and inconsistent recognitions (on a scale of 0 to 3) in both contexts, with the standard error, are displayed in panel (b).

3.4 CONCLUSION

In the first part of this chapter we gave an overview of the knowledge about human perception of sound events, which is acquired through studies in multiple research areas, such as psychoacoustics and neuroscience. Although the presented overview is not exhaustive, it can provide a basis for automatic sound event recognition. First, a model for automatic sound event recognition can be guided by the properties of human perception (see section 3.2.2 and Table 3.1). These properties can be summarized by two attributes, separability and constancy, which inspire our design of a model for automatic sound event recognition. A system that should be robust to a changing environment benefits from a representation of the input that is constant. In other words, sound events need to be separated from the background, and they should be stored (remembered) as invariant representations in our model.

Second, different abstraction levels in a model for automatic sound event recognition correspond to different levels of precision in the analysis of sound events, depending on the goal or mode of a system. A high level of precision is difficult to obtain, but may be possible in a known and controlled environment, for example in automatic speech recognition systems that work with a close-talking microphone. In these applications, statistical models based on Bayes theorem can calculate probabilities of sequences of phonemes, and transcribe spoken words and sentences with high accuracy. However, in a real-world environment with many unknown and variable events, it is difficult (if not impossible) to determine exact probabilities. Hence, statistical decision models are unsuitable for these situations (Box 3.3). Analogous to human perception, the problem of recognizing sound events gets easier when it is described at a higher abstraction level, although the precision will be lower. Therefore, we choose to design our model such that it relies on more rough estimations instead of exact probabilities. As a consequence, it should be more robust to unknown and variable conditions.

In the second part of this chapter we investigated the effect of context (in the form of a sound event preceding a target sound event) on the recognition of homonymous sound events by people. We confirmed the consensus in vision research that context can facilitate the recognition of an object or event. This result is valid within this experimental setup, but has to be further explored in different

Box 3.3: Decision processes

Decision processes in speech recognition systems are made based on conditional dependencies between models of phonemes (or other speech elements), taking a statistical language model into account. Typically, these systems work with feature vectors that describe the spectrum of the audio signal 100 times per second (see Figure 1.2). The resulting high-dimensional feature vectors are used in a probability multiplication process, in which the probability of a long pronunciation of a word, for example "heeeeeelp", will be lower than for a normal pronunciation ("help"). As a consequence, alternative interpretations of an erroneous sequence of words of normal duration can be favored over the actual utterance. In contrast, people would note only a duration difference between the two utterances.

experimental designs. For example, in some other design a mismatching context can facilitate recognition, because it makes a target stand out (Gygi and Shafiro, 2006). However, automatic sound event recognition can benefit from the facilitatory effect of context when the audio signal is ambiguous. Hence, this effect can be modeled to improved automatic sound event recognition, provided that the contextual information obeys the form in which it is shown to enhance recognition in human perception.

4

Modeling context of sound events

The content of this chapter has been published as part of Niessen, M. E., Van Maanen, L., & Andringa, T. C. (2008). Disambiguating sound through context. International Journal on Semantic Computing 2(3), 327–341. Proceedings of the 2nd IEEE International Conference on Semantic Computing, 88–95.

A central problem in automatic sound recognition is the mapping between signal-driven audio patterns and the semantic interpretation of a sonic environment. We propose a context model to perform this mapping. Acoustics research is predominantly devoted to mimic early stage human perceptual abilities such as audio pattern selection and grouping, which are translated into successful signal processing techniques. In contrast, not many studies are aimed at modeling knowledge and context in sound recognition, although this information is necessary to recognize a sound event in addition to segregating its components from a scene. Based on the investigation of the role of context in human sound event recognition in the previous chapter, we show that the use of knowledge in a context model can improve automatic sound event recognition by reducing the search space of the signal-driven audio patterns. Furthermore, context information dissolves ambiguities that arise from multiple interpretations of one sound event.

4.1 INTRODUCTION

In acoustics much research is devoted to modeling the ability of the human auditory system to segregate different events in a sonic environment based on the audio signal alone, called primitive auditory scene analysis (ASA, Bregman, 1990). Perceptual grouping based on features such as continuity of components in the audio signal and proximity in time or frequency are translated into successful models of primitive ASA (Cooke and Ellis, 2001; Godsmark and Brown, 1999; Grossberg *et al.*, 2004; Nix and Hohmann, 2007; Wang and Brown, 2006). However, primitive ASA alone will not suffice to automatically recognize sound events. We also need to model the contribution of knowledge and context to interpret the audio signal and make predictions at a higher description level (see chapter 3). Although this need has been recognized some time ago (Ellis, 1996, 1999), it has so far not resulted in models of sound event recognition that combine signal-driven (bottom-up) and *context-based* (top-down) methods.

In recent years there has been some research on modeling what is called context awareness in sound recognition. One group of studies focusses on estimating the context of an audio interval with varying classification techniques (Eronen *et al.*, 2006; Chu *et al.*, 2009; Aucouturier *et al.*, 2007). In these studies the context is represented by a class of sounds that can be heard in some type of environment, such as cars at a street, or people talking in a restaurant. Depending on the number of context classes that are learned, the recognition rates of these methods vary between 58% (24 classes, Eronen *et al.*, 2006) and 84% (14 classes, Chu *et al.*, 2009). Although these results are promising, the methods that are used have some attributes that make them less suitable for automatic sound event recognition.

First, no sound event segregation (see section 2.3) is applied, so the features that are used to classify an audio interval are assumed to represent information that is specific for a class. Therefore, the context class to which an audio interval belongs gives primarily information about its acoustic properties. Consequently, these methods implicitly assume that similar sound types occur in similar situations. Although this assumption is valid for some tasks in a restricted domain, like music genre determination, it cannot be guaranteed for sound events in real-world environments. The context (as defined in section 3.1) of an event in a real-world situation is not only the acoustic class it may be categorized into, but the envi-

ronment in which it can be heard as well. For example, speech can be heard in restaurant, together with music and clinking of glasses, but also in a park, together with birds and wind.

Second, tasks in multimedia applications (or a comparable setup in environmental sound classification, as in Chu *et al.*, 2009) generally entail that a small audio interval, in the studies above typically not longer than a few seconds, is classified as a sample of one context out of a data set with a limited set of distinct contexts, which are stored as a collection of audio files. The estimation of a stationary context is easier when the intervals are longer, because the reliability of the audio features increases with time, assuming the audio signal does not change qualitatively. In contrast, in a real-world environment it cannot be assumed that the sonic composition within a context—where context refers not to the type of sound, but to the associations between events and environments (see Box 3.2 on page 35)—does not change over time. Furthermore, in continuous monitoring of a sonic environment there is no prior segmentation of interesting sound intervals. Therefore, the information in the sound used to define the context is not necessarily relevant, as it is assumed to be in the audio files in multimedia applications. To be able to determine the environmental context of a sound event, the sound event needs to be segregated and recognized, and co-occurring events that are semantically related to the same context can help to estimate the most likely context.

A second group of studies on context awareness addresses some the above issues by retrieving semantic relatedness of sound intervals rather than the similarity of their acoustic properties (Lu and Hanjalic, 2008, 2009; Cai *et al.*, 2006, 2008). For example, in Cai *et al.* (2006, 2008) the intervals are clustered based on the similarity of their audio features. Subsequently, the semantic relatedness of these intervals to each context (different tracks from shows on television, such as series or tennis games) is calculated based on their co-occurrences. However, the audio intervals that are selected give no information about the semantics of the interval itself, only about the context it belongs to. Since we want to recognize events in a real-world sonic environment, we are not interested in the context per se, but in its usage to improve recognition.

Furthermore, sound event recognition is different from segmentation, where an audio signal is divided into different intervals based on features, such as the zero crossing rate in the time domain and the spectral centroid, that represent prop-

erties of the signal (Tzanetakis and Cook, 1999). Figure 4.1 displays a schematic overview of the segmentation process. The segmented intervals do not necessarily correspond to events. Sound events in a real-world environment are often heard simultaneously, or no (interesting) events are heard at all. In other words, we do not want to know the acoustic class of each interval, but which events are present in a continuous stream of audio. Context can help to limit the search space of possible sound events.

While context-based recognition has received little attention in automatic sound recognition—with the notable exception of speech, where grammatical and lexical rules are considered for automatic recognition (Barker *et al.*, 2005; Scharenborg, 2007)—it has a long history in other research areas such as information retrieval (Cohen and Kjeldsen, 1987; Crestani, 1997; Van Maanen, 2007; Van Maanen *et al.*, 2010) and handwriting recognition (Côté *et al.*, 1998; McClelland and Rumelhart, 1981). Models of context-based recognition assume that certain regularities exist in the contexts in which an event may occur and structure their knowledge base in such a way that these regularities are accounted for. Often this takes the form of a spreading activation semantic network (Quillian, 1968; Collins and Loftus, 1975), in which the nodes represent the states the network can be in, and the vertices represent the prior probabilities that these states are encountered subsequently or together. In these models, context is incorporated by keeping nodes active over a longer period of time, thereby influencing the probabilities that certain nodes will be activated. Spreading activation networks have mostly been exploited in static and well-constrained domains. Our aim is to demonstrate that spreading activation can also be applied in a dynamic domain such as sound event recognition.

4.2 CONTEXT MODEL

Based on existing models of spreading activation and the findings of the context facilitation experiment in section 3.3 we introduce a model for context-based recognition that can be used with dynamic real-world audio input. This model allows automatic recognition of events in a complex and changing sonic environment. In complex real-world environments a sound event may have different interpretations, depending on the situation in which it occurs. Therefore, the model needs knowledge about the context to interpret the audio features, similar to humans.

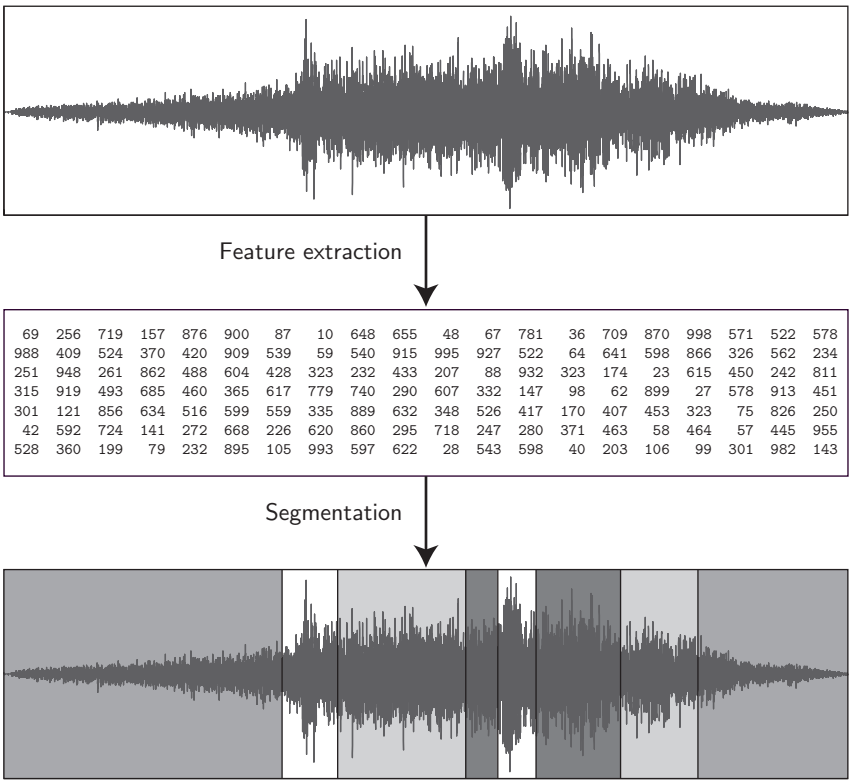


Figure 4.1: Schematic overview of a segmentation process. Acoustic features, such as the zero crossing rate and the spectral centroid, are extracted from the audio signal. Based on the feature values for the different time frames, the audio signal can be segmented into different intervals that have corresponding feature values, indicated by gray scales in the bottom panel.

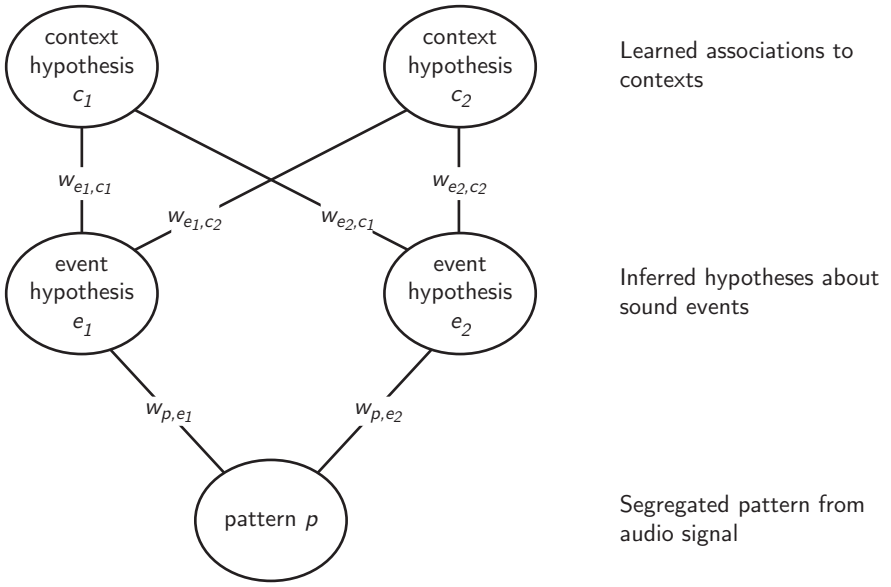


Figure 4.2: Schematic overview of a network initiated by a pattern segregated from an audio signal. The pattern is connected to event hypotheses that are possible interpretations of the pattern. The learned associations to contexts can help to infer the most likely interpretation for a certain pattern. All possible connections are depicted as lines. The strength of these connections is indicated with a symbol w.

The model dynamically builds a network that generates semantic hypotheses of sound events based on signal-driven audio patterns and knowledge of the events. Moreover, context information is used to compute the support for competing hypotheses, and consequently a most likely hypothesis for all segregated patterns can be assessed. Figure 4.2 shows a schematic representation of a network.

With our model we want to qualitatively improve automatic sound event recognition. Our approach starts with signal-driven techniques for the selection and grouping of audio components (the methods for segregation are explained in section 5.2.1). Every segregated pattern represents a possible sound event. The ability of people to use their knowledge of the context to disambiguate sounds, which we demonstrated in the experiment in chapter 3, should also be present in the model. Therefore, a context model evaluates the signal-driven input with knowledge of

Table 4.1: Algorithm for updating the dynamic network configuration at times when new signal-driven information is presented to the network.

For all patterns (grouped components) at time t:	
1.	Segregate the audio signal into patterns $P(t) = \{p_{t,i}\}$
2.	For each $p_{t,i} \in P(t)$ add possible event hypotheses $\{e_{t,j}\} \in E(t)$ and connect them with strength $w_{i,j}$
3.	For each new event $e_{t,j} \in E(t)$ add appropriate contexts c_k not yet present in the set of active contexts C
4.	Connect each new event $e_{t,j}$ with the appropriate context c_k with strength $w_{j,k}$
5.	Spread signal-driven activation
6.	Spread context-based activation
7.	Evaluate activation values

the event and its context (Andringa and Niessen, 2006). We will illustrate the behavior of the model through an example of a sound event that was also used in the experiment, the mix of a bouncing basketball and a closing door, which can be identified as both in the absence of context information. In the following sections we will describe how the model dissolves this ambiguity through the use of context knowledge in a dynamic network. The algorithm for the construction and updating of the dynamic network is summarized in Table 4.1.

4.2.1 Dynamics of context model

Because we want to combine a signal-driven (bottom-up) and context-based (top-down) approach to sound recognition, the model maps hypotheses of the sound event based onto segregated audio patterns to expectations that are formed by knowledge of the relations between the events and the context. This mapping process will lead to a best hypothesis about the event that causes the sound in this context, at every description level in the network (apart from the lowest, which are the segregated audio patterns). All hypotheses hold a confidence value reflecting their support from relations to other events and the context in which the hypothesized event is occurring. In case of conflicting interpretations for one event, the hypothesis with the highest support will win. For example, in Figure 4.3 the reverberant impact sound could be either a closing door or a basketball bouncing,

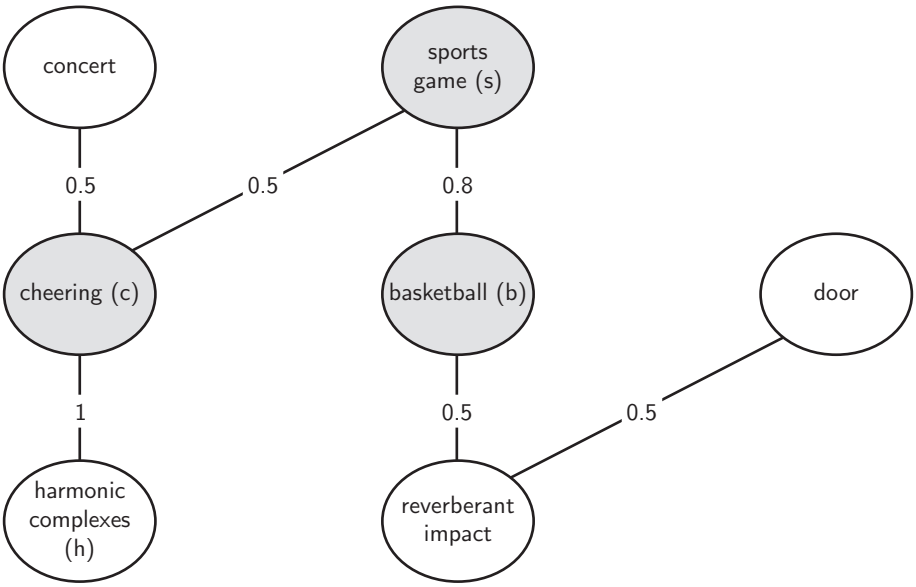


Figure 4.3: Network configuration for the identification of a reverberant impact sound preceded by the sound of cheering. The best hypotheses at the highest two levels (the gray nodes) correspond to the best interpretation for the signal-driven evidence at that description level.

based on the audio pattern alone. However, knowledge about the context actuated by a previous sound event (cheering) will increase the support for the hypothesis that the second sound event is a basketball bouncing. Furthermore, the confidence value of the first hypothesis (cheering) is increased, because the context of a sports game, and hence the cheering, is more likely considering the new input. In the following paragraphs we will describe the process of how the network is dynamically built, and how the confidence of all hypotheses is established through spreading activation.

The network is updated if and only if new signal-driven information is presented, and spreads its activation when the network is stable, that is, when the available knowledge about the signal-driven information is processed. The hierarchy in the network is captured by the interdependent relations of all the hypotheses. The lowest description level in the network corresponds to the physics of the

signal, and the highest level to a (provisional) interpretation of the environment. The intermediate levels represent hypotheses of increasing generality. The number of levels depends on the complexity of the domain, but usually three levels will suffice: one for the segregated patterns, one for the event hypotheses that are inferred from the patterns, and one for the context of the environment, which can raise particular expectations about future events (Figure 4.2).

4.2.2 Algorithm of context model

In the first step (see Table 4.1), audio patterns that are likely to be caused by sound events are segregated from the time-frequency plane of the sound. Figure 4.4 shows the time-frequency plane of a sports game scene with annotated audio patterns.[1] Every hypothesis of a sound event corresponds to a specific pattern of audio components. For example, the cheering is a noisy collection of distorted harmonic complexes. Each segregated event comes with a base-level activation based on the confidence given by the segregation algorithms. For example, a confidence value may reflect how well a pattern, such as a harmonic complex, fits a particular mask, such as a calculated harmonic complex. For illustrative purposes we set the base-level activations of all patterns to 1 in the example.

A segregated pattern of audio components may have multiple interpretations. Hence, all proposed interpretations of a pattern will be initialized as sound event hypotheses (step 2 in Table 4.1). These event hypotheses are connected to the patterns that initiated them, with some strength, denoted by weight w. Subsequently, knowledge about the hypothesized events will initiate hypotheses about the context in which an event can occur, for example about an event sequence or an environmental setting[2] (step 3), and connected to the event hypotheses (step 4). In Figure 4.3, the cheering could point to a pop concert or a sports game. These higher level context hypotheses create expectations about sound events that will follow, like a basketball in a sports game. If the expected event is matched with signal-driven evidence, it will receive extra support when its hypothesis is created.

[1] In the application of the model we use grouped audio components that are automatically extracted from the audio signal instead of annations (see chapter 5).

[2] The knowledge about the context is learned in a training phase. The method used for training is discussed for each application, that is, section 5.2.2 and 6.2.2.

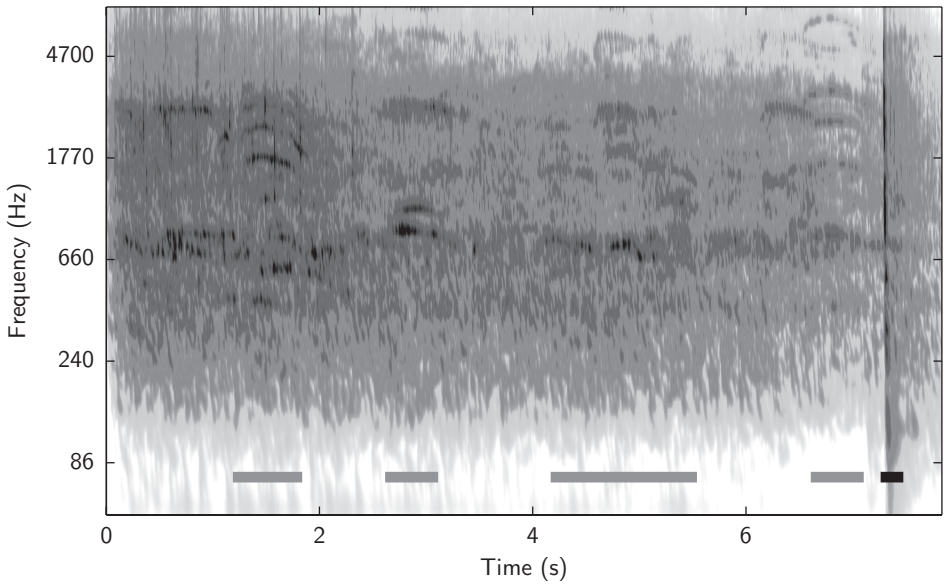


Figure 4.4: The cochleogram of the sound of cheering followed by a chimaeric basketball/door sound. The gray bars indicate annotations of harmonic complexes and the black bar the annotation of a reverberant pulse.

When the knowledge is processed and the network is stable, the activation of the audio pattern spreads through the network (step 5 and 6).

The connections in the dynamic network are symmetrical, and only between hypotheses at different levels. For instance, the event hypotheses are connected to the segregated patterns that initiated them, and to hypotheses of their possible contexts, but not to each other (Figure 4.2). Connections between hypotheses at the same level would be redundant, since they can reinforce each other through shared parent hypotheses. Furthermore, the hierarchy of the network is now captured by the connections. Therefore, it is not necessary to store a global representation of the complete network. Instead, each hypothesis contains information of its relative position in the network, that is, it stores its direct connections. The only information that is globally available is which hypotheses are active.

4.3 ACTIVATION SPREADING

When the network configuration is stable after updating, the activation first spreads upward to the highest level in the network (signal-driven activation spreading, step 5), and then downward to other connected events in the past, if they exist (context-based spreading, step 6). The spreading can only go up once and down once through every path that denotes a past event, after which it terminates. The activation of the individual hypotheses is a time-dependent weighted sum that decays exponentially with time. The rate of decay is determined by a time constant τ. The activation of each hypothesis is limited to a maximum value. As a consequence, hypotheses that are highly active over a longer period of time are not repeatedly reinforced by new input, because the effect of the input decreases when the activation of a hypothesis reaches its maximum. Because the connection strength between the hypotheses are learned on training data (section 5.2.2 and 6.2.2), the activation represents a pseudo-probability (confidence) that the hypothesis is true.

The computation of the spreading activation is similar to the method used in the model of letter perception by McClelland and Rumelhart (1981). However, we only incorporate excitatory and no inhibitory connections. Furthermore, the decay function applied in our model is a continuous function of time instead of a constant value that is applied at discrete time steps.

The input activation $n_i(t)$ of the individual hypotheses is the weighted sum of all connected hypotheses, either from the level below, for signal-driven activation spreading (step 5), or from the level above, in case of context-based spreading (step 6):

$$n_i(t) = \sum_j w_{ji} A_j(t), \tag{4.1}$$

where j is a hypothesis connected to i, $A_j(t)$ is its activation, and w_{ji} is the connection strength between hypotheses j and i, retrieved from the stored knowledge.

When an audio pattern holds a low confidence value, the activation spreading from the higher levels to the event hypotheses is more important than the activation spreading upward from the pattern, and vice versa. In other words, the lower the saliency of the signal, the more influential the context is. As a consequence, the dynamic network is more robust to unreliable input than models that rely only on signal-driven techniques.

4.3.1 Activation evaluation

After the activation has spread through the network, the activation of each hypothesis is evaluated (step 7). The activation evaluation is an accumulation of current input and the previous activation corrected with a decay. The decay ensures that items in short-term memory are forgotten without reinforcement (new signal-driven evidence) in contrast to information in long-term memory (Quillian, 1968). The activations of all hypotheses decay exponentially with time toward a default situation. Therefore, the decay function is dependent on the a priori activation of a hypothesis:

$$f_i(\Delta t) = e^{-\frac{\Delta t}{\tau}}(1-\tilde{A}_i)+\tilde{A}_i, \tag{4.2}$$

where $\tilde{A}_i$ is the default activation of hypothesis i, which is non-zero for a closed set of contexts. For example, when the context is represented by one of N locations, the default activation can be determined from their incidences in the training data. In such a situation, the sum of the default activations of all context hypotheses is 1. For all other hypotheses $\tilde{A}_i = 0$. Furthermore, τ is a time constant controlling the rate of decay, and Δt is the elapsed time since hypothesis i is evaluated last. As a result, hypotheses deactivate when they do not receive input activation from other hypotheses. When the activation value decreases below a minimum value (θ_A), the hypothesis is no longer evaluated, and removed from the dynamic network. A new hypothesis will be initiated when new evidence is found for the same type of event.

The activation value of the hypotheses is normalized to the maximum input activation, so that it is scaled between 0 and 1:

$$A_i(t) = f_i(\Delta t)A_i(t-\Delta t) + n_i(t)(M - f_i(\Delta t)A_i(t-\Delta t)), \tag{4.3}$$

where $M = 1$ is the maximum activation level, and $A_i(t-\Delta t)$ is the activation of hypothesis i when the network was last updated, multiplied with a decay $f_i(\Delta t)$, computed according to equation 4.2. Furthermore, $n_i(t)$ is the input activation as calculated in equation 4.1. It should be noted that the activation of the audio patterns will always decay, because they do not get any more input activation ($n_i(t) = 0$) after being initiated. In contrast, event and context hypotheses can get reinforced by new evidence from subsequent audio patterns, and thus can stay active for a longer period of time. The result of the activation evaluation of a hypothesis is treated as the pseudo-probability that the hypothesis is true.

Going back to the example of Figure 4.3, the activation of the sports game hypothesis is summed over the two time steps when new signal-driven information is presented to the network. The value of the time constant τ is arbitrarily set to 100 to demonstrate its effect in the calculation of the activation value. However, in different application domains the value of τ can be estimated based on training data. At the first moment, the activation of the sports game hypothesis consists of the input it gets from the cheering hypothesis, which starts at time $t = 1.7$ (the subscript letters are in parentheses in Figure 4.3):

$$A_s(1.7) = n_s(1.7) = w_{cs}A_c(1.7) = 0.5 * 1 = 0.5 \tag{4.4}$$

A few seconds later, at time $t = 7.7$, the input is delivered by the basketball hypothesis:

$$\begin{aligned} A_s(7.7) &= e^{-\frac{7.7-1.7}{100}}A_s(1.7) + w_{bs}A_b(7.7)(1 - e^{-\frac{7.7-1.7}{100}}A_s(1.7)) \\ &= 0.94 * 0.5 + 0.8 * 0.7 * (1 - 0.94 * 0.5) = 0.77 \end{aligned} \tag{4.5}$$

The activation of the cheering hypothesis is not included in equation 4.5, because at every update only the active connected hypotheses can deliver input to the sports game hypothesis. As a consequence of the two-way spreading, the cheering hypothesis will receive an increased support from the basketball bouncing, through the sports game hypothesis. In the first step the hypothesis receives activation from the signal-driven evidence:

$$A_c(1.7) = n_c(1.7) = w_{hc}A_h(1) = 1 * 1 = 1 \tag{4.6}$$

In the second step the sports game hypothesis contributes to the activation of the cheering hypothesis:

$$\begin{aligned} A_c(7.7) &= e^{-\frac{7.7-1.7}{100}}A_c(1.7) + w_{sc}A_s(7.7)(1 - e^{-\frac{7.7-1.7}{100}}A_c(1.7)) \\ &= 0.94 * 1 + 0.5 * 0.77 * (1 - 0.94 * 1) = 0.96 \end{aligned} \tag{4.7}$$

The activation values of all the higher level hypotheses are shown in Figure 4.5.

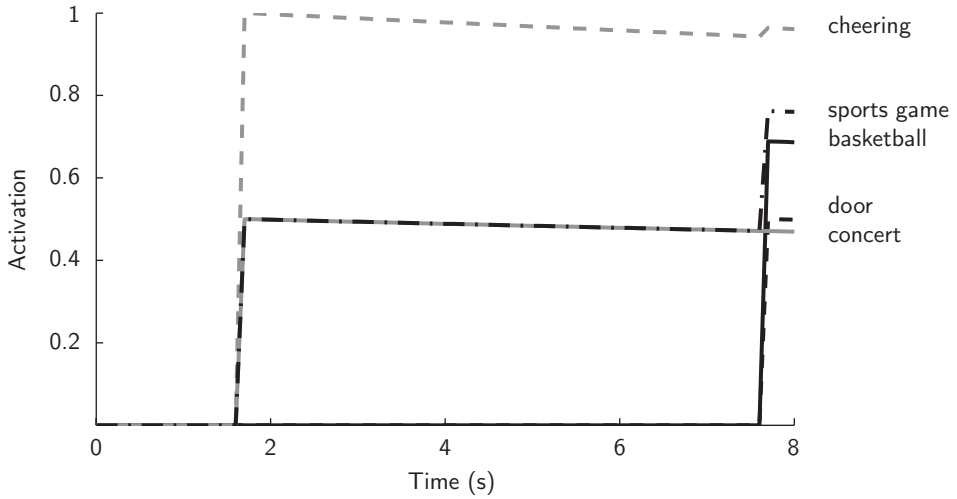


Figure 4.5: Activation of all higher-level hypotheses in the example of Figure 4.3. At the two moments when signal-driven patterns are presented to the network, the activation values of the hypotheses that are connected to these patterns increase.

4.4 CONCLUSION

The network described in the example is rather simple, while in a real-world environment there will be many more events, mostly of unreliable sound quality. The complexity of a real-world environment will be captured by knowledge about the relations between the real-world events. Furthermore, the expansion of the network must be controlled. This is partly achieved by keeping track of which hypotheses are active, and which hypotheses are of finished or discarded events. These last two classes are excluded from the search space of connected hypotheses when new information is presented to the network. As a consequence, the search space at any time is limited to the hypotheses that are active at that time. An advantage of a complex environment is its supply of information. People use much more contextual information in the recognition of sound events, such as time of day, environmental setting and ecological frequency (Ballas, 1993). This information can also be included in our model in the form of nodes in the network that help support or discard hypotheses.

In the next part we will show applications of the model in real-world situations,

where the input patterns of the model are supplied by automatic audio signal segregation algorithms. Furthermore, we will show how the knowledge of the model is acquired in the training phase of an application. Although the model is being developed for audio input, its general implementation allows for other signal-driven input, such as image descriptions, as long as they represent a single event or object (chapter 6). If different types of descriptions can serve as input to the model, they may be combined in one model for use in multimedia applications.

5

Automatic sound event recognition in the real world

The content of this chapter (except section 5.3) has been published as Niessen, M. E., Krijnders, J. D., & Andringa, T. C. (2009). Understanding a soundscape through its components. In Proceedings of Euronoise.

The content of section 5.3 has been published as part of Krijnders, J. D., Niessen, M. E., & Andringa, T. C. (2010). Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. Pattern Recognition Letters 31(12), 1552–1559.

Human evaluation of sound in real environments is a complex interaction of many factors, which are investigated by a range of research fields. Most approaches to assess and improve the evaluation of sonic environments (soundscapes) use a holistic approach. For example, in environmental psychology, subjective measurements usually involve the judgment of a complete soundscape by people, mostly through questionnaires. In contrast, psychoacoustic measurements try to mimic perceptual attributes by measurements such as loudness. However, these two types of soundscape measurements are aimed at qualitatively different phenomena, which are difficult to link other than with correlational measures. We propose a method grounded in cognitive research to improve our understanding of the link between sound events and human soundscape evaluation. People process sound as meaningful events. Therefore, we developed a model to recognize sound events in a sonic environment that are the basis of these meaningful events.

5.1 INTRODUCTION

Human evaluation of sound in real environments is a complex interaction of many factors. Therefore, many fields of research are involved in trying to understand these factors, ranging from psychoacoustics (Fastl, 1997) to cognitive psychology (Guastavino, 2007) and sociology (Schulte-Fortkamp and Fiebig, 2006). Most psychoacoustic studies on sound quality evaluation focus on measuring attributes such as loudness and sharpness of isolated sounds (Blauert and Jekosch, 1997; Fastl, 1997). However, several studies have shown that these perceptual attributes can only explain part people's evaluation of soundscapes (Ballas, 1993; Maris *et al.*, 2007). The judgment of a soundscape largely depends on the meaning that a person gives to the sound (Dubois *et al.*, 2004). For example, whether a person enjoys music depends on his or her choice to hear it. At a concert, music will be appreciated even (or especially) at a high loudness level, while the tolerance for the music that a neighbor is playing at night will be much lower.

Zhang and Kang (2007) distinguish four categories in which the different factors that influence soundscape evaluation can be organized, namely sound, space, people, and environment. The categories of sound and space comprise the acoustic factors of the sound events in the environment, modified by transmission effects, such as background sounds and reverberation (see chapter 2). These acoustic factors in soundscape evaluation have been tested through psychoacoustic measurements (Genuit and Fiebig, 2006) and questionnaires (Raimbault *et al.*, 2003), amongst others. However, acoustic factors cannot be studied in isolation, because the judgment of people is not based solely on the properties of the sound, but is also affected by the meaning they give to the environment and to the sound events in the environment. This meaning depends on factors like their memory and cultural background. The interplay between acoustic and non-acoustic factors can be examined either by controlling the sounds and varying the condition, such as the cultural background (Hansen and Weber, 2009), or by correlating psychoacoustic measures to people's judgments (Raimbault *et al.*, 2003). Other non-acoustic factors that affect people's judgments can include social, demographical, and behavioral factors (Yu and Kang, 2008), and environmental factors, such as temperature, wind and sunshine (Zhang and Kang, 2007).

Many of these recent studies on the judgment of soundscapes use measure-

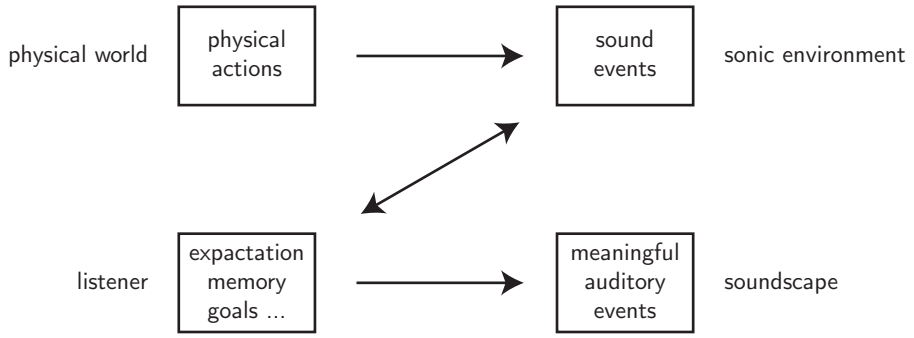


Figure 5.1: Schematic overview of the link between sound events and a soundscape. Physical actions result in sound events, which are interpreted by a listener based on his expectation, goals, and so forth. We refer to a sonic environment as the collection of sound events, and to a soundscape as the collection of meaningful auditory events. (The image is not meant to be exhaustive. For example, the effect of visual information is not included.)

ments that describe the sonic environment as a whole, and not the individual sound events. However, people judge a sonic environment not (only) based on the acoustic properties of the sonic environment. Rather, they evaluate the soundscape by the meaning they give to the environment and the sound events (Guastavino, 2007). For example, Guastavino (2006) showed that an important indicator for the judgment of an urban soundscape is the presence or absence of human activity. Whether there is human activity can be determined by sound events that result from the presence of people, such as speech. Therefore, meaningful events that indicate either a pleasant or an unpleasant situation can function as a link between the holistic judgment of a soundscape and the sound events that constitute the sonic environment (Figure 5.1).

Because sound events are indicators for soundscape evaluation, we propose a method to automatically recognize sound events based on signal-driven hypotheses, which are guided by knowledge of the environment. The recent history of a sound event is used to estimate the context (Box 3.2, page 35) where the event is recorded. Subsequently, the predicted context is used to form expectancies of future events. The hypotheses about events are approximations of meaningful events, because they are learned from human annotations. Furthermore, we use a model

of human memory to represent the hypotheses, through which we include an important cognitive factor in the analysis of a soundscape. Although these event hypotheses are not similar, or even close to human cognitive representations, they can be used to automatically analyze a soundscape based on knowledge other than acoustics. Therefore, this method provides a basis for modeling the factors that are important in soundscape evaluation.

In the next section we will describe the methods that we developed to segregate and interpret sound events. In the third and fourth section we present two experiments with two different data sets to test the integrated methods on their comparison to human annotations. Finally we will discuss the results of the experiments, and provide an outlook on future work.

5.2 METHODS

To recognize sound events in a continuous audio signal, we first segregate patterns from the audio signal that are likely to constitute a single event. Subsequently, we use a model of human memory to select the most likely interpretation for the events that these patterns represent, based on an estimation of the context.

5.2.1 Audio processing

The cochleogram of the audio signal is segmented on the basis of the local spectro-temporal properties. Segments are likely to be produces by a single event when they are based on local properties. For example, local energy maxima that resemble tones and are developing smoothly in time are likely to be produced by the same event. 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). Figure 5.2 shows an example of a cochleogram of a speech signal, of which the harmonic components are selected and grouped. The strategy to combine local signal properties and grouping principles allows one to select qualitatively different types of audio patterns, namely tones and harmonic complexes, pulses, and broadband events (see appendix B and Krijnders, 2010). A description of these patterns based on their signal properties is used to classify and label them as sound events with a machine learning algorithm.

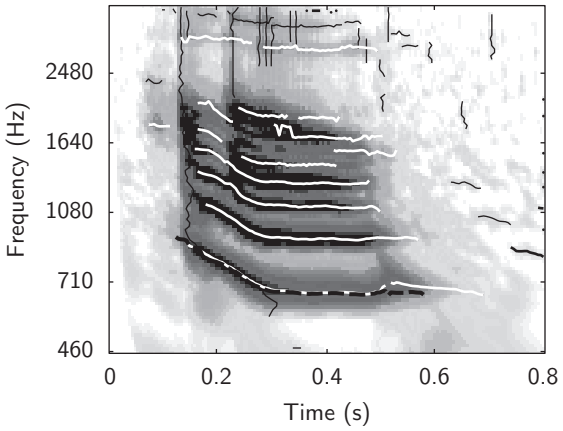


Figure 5.2: Cochleogram of a speech signal with segregated signal components (black lines). The white lines are the components that have been grouped into a harmonic complex, of which the fundamental frequency is depicted by the dashed black line. The vertical black line indicates the onset of the speech.

5.2.2 Context model

The segregated patterns, described in the previous section, are classified with a standard machine learning technique. However, the audio signal can be distorted or masked by transmission effects such as background sounds and reverberation, resulting in a low confidence of certain pattern labels. Furthermore, some sound events can have multiple interpretations while they share similar audio patterns, such as screaming and laughing. To find the most likely interpretation, we introduce a method that incorporates knowledge of the context.

This method, which is inspired by research in cognitive psychology (Quillian, 1968; McClelland and Rumelhart, 1981), constructs a dynamic network that keeps track of both signal-driven patterns and knowledge of the context (see chapter 4). The nodes of the dynamic network represent information about sound events at different description levels, and the vertices between them represent the probability that these pieces of information belong together. Each node holds an activation value. A hypothesis (node) is more likely to represent a relevant interpretation when its activation value is higher than its competitors. Whenever new signal-driven information becomes available, the network is updated by adding nodes,

which each represent new pieces of information, and removing nodes whose activation decreased below a threshold. Subsequently, the activation of the new nodes spreads through the network. Furthermore, new nodes are used to form expectancies of future sound events. If a signal-driven pattern matches an expected event, it is more likely to be an adequate interpretation.

Knowledge network

An example of a network configuration is depicted in Figure 4.3. The nodes at the lowest level represent segregated audio patterns, at the middle level they correspond to sound event hypotheses, and at the highest level to hypotheses about recording locations or sequences of co-occurring sound events. Nodes at the different levels are connected with some strength, denoted by weight w. These weights are learned in the training phase and stored in a knowledge network. In the operation phase the weights are used to infer a probable context of sound events. In the experiments presented in this chapter the context can either refer to a sequence of events (section 5.3) or a recording location (section 5.4).

The strength between the node that represents a recording location or a sequence and the nodes that represent the 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 (term frequency) with its general frequency in other documents (inverse document frequency). 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 e is encountered often at recording location l, and little at other locations, it is an important indicator for location l. Accordingly, the strength between the sound event e and the recording location l is:

$$w_{l,e} = w_{e,l} = \mathrm{ef} \cdot \frac{\log_{10} N - \log_{10} n}{\log_{10} N} \tag{5.1}$$

where N is the total number of recording locations, n is the number of locations at which e occurs, and the term (that is, event) frequency is given by:

$$\mathrm{tf} = \frac{T_{e,l}}{T_e}, \tag{5.2}$$

where $T_{e,l}$ is the total duration of occurrences of e in l, and T_e is the total duration of occurrences of e in a training set.

The term frequency of events is calculated with the duration of events instead of their frequency, which is common in automatic document retrieval, because duration is more robust to variations in annotations of different human annotators and of different files (see the data descriptions in section 5.3 and 5.4). For example, some annotators choose to annotate every single bird, while others annotate a complete file as containing bird sounds. Therefore, the difference in the number of annotations can be considerable compared to the difference in the duration that is annotated. In other words, what constitutes one instance of an event is more difficult to judge than whether the event is present.

Furthermore, the inverse document frequency (IDF, the second part of equation 5.1) is normalized, such that a sound event that occurs at one location (or is part of one sequence) has an IDF of 1, while a sound event that is recorded at all locations has an IDF of 0, regardless of the frequency it is recorded at those locations (the term frequency). In other words, a sound event that can be heard everywhere does not provide any information about the location of a recording.

Equation 5.1 can also used to determine the weights between sound events and sound sequences, instead of recording locations. If a sound event A is encountered often in combination with some sound event B, and infrequently with other sound events, it is important in the event sequence $s : A-B$.[1] In this case, l in equation 5.1 can be replaced with s, which represents the event sequence instead of the location, so $w_{e,l}$ is substituted with $w_{e,s}$. Additionally, N is the total number of sequences, and n is the number of sequences in which event A of sequence s occurs.

Whether the knowledge network is trained on recording locations or sequences of events can be decided based on the data set. In a data set with recordings at different types of locations, the location can be predictive of the sound event, while this information would be useless in a data set collected at a single location. In such a data set it is more sensible to use location independent information to predict sound events, for example, which sound events co-occur, or at what time they occur.

[1] Two events are considered to comprise a sequence when they co-occur within a certain time frame.

Activation evaluation of event sequences

The spreading of the activation through the network, and the evaluation of the resulting activation values of the hypotheses in the network is determined according to the algorithms that are explained in section 4.3.1. However, the activation evaluation of the event sequences with a fixed order is extended compared to the other hypotheses. 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. Like all other hypotheses, the expected activation of an event that is part of a non-fixed event sequence is calculated by multiplying the activation value of the event sequence with the connection strength between the sequence and the type of event. Since the activation value decays with time, the expected value is smaller when the other event of the sequence occurred longer ago.

However, for some sequences the order can be very indicative. For instance, in the data set of the first experiment (section 5.3) there are trains departing, which are normally preceded by a whistle of the conductor. Hence, if a whistle is identified, a strong expectancy of a departing train should arise. To capture the regularity of ordered sequences, we determine whether the sound events that constitute a sequence have a strong bias to a specific order. For these ordered sequences, the first sound event primes the network for the second sound event after a time interval learned from examples in the training data. The mean time difference between the events is used in a function to calculate the expected activation value of the second event in the sequence:

$$\hat{A}_i(t) = w_{ji} A_j(t - \Delta t) \mathrm{e}^{\frac{-(\Delta t - \overline{T})^2}{2\sigma^2}}, \tag{5.3}$$

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

Instead of applying a decay to the primed sound event hypothesis (equation 4.3), its activation is determined by weighting the signal-driven evidence with the

expected activation value:

$$A_i(t) = \hat{A}_i(t) + K\left(\frac{n_i(t)}{\max n(t)} - \hat{A}_i(t)\right), \tag{5.4}$$

where $\hat{A}_i(t)$ is the expected activation according to equation 5.3, $n_i(t)$ is the input activation of i as calculated in equation 4.1, $n(t)$ is a list with the input activations of all active sound event hypotheses, and K is the gain factor. The gain factor is dependent on the complexity of the audio signal. If the segregated patterns are very salient, its value should be high. However, sound recorded in real-world environments, as in the presented experiments, are relatively noisy. Therefore, the gain factor in the first experiment is set to 0.5, which entails that the model responds relatively slowly to new evidence from the signal, and is guided equally much by expectancies.

5.3 EXPERIMENT 1

The purpose of this experiment is to demonstrate that the proposed methods can be used to improve the recognition of sound events in a rich outdoor environment, using knowledge of co-occurring events. The data set used in this experiment has been created to develop and test aggression detection systems (Zajdel *et al.*, 2007), and is recorded on a busy train station. Therefore, it includes problems of real-world environments, such as unknown transmission effects and ambiguous sound events. For example, the sound of a train and a subway train are very similar. Based on the audio signal alone, humans have problems recognizing such events as well. In the next section we describe the data set that is used in the experiment. Subsequently, the setup of the experiment is explained, and in the last part we present the results.

5.3.1 Data

The data set consists of 40 enacted scenes from 16 different scenarios, which last between one and two minutes each (Zajdel *et al.*, 2007). The total duration of the recordings is 54 minutes. The scenes were acted by professional actors (three men and one woman) on a platform of the station Amsterdam Amstel. 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, such that all scenarios were played out

twice or more with different actors. The 16 scenarios were based on events that are likely to happen at stations, like friends meeting, shouting football supporters, and diverse forms of verbal aggression and vandalism. The scenes were recorded with 8 microphones (16 bits, 44.1 kHz sampling rate), of which one was used for this experiment. This microphone was located about two meters from the center of the action and about two meters from the subway track. The scenes were also captured by three cameras.

The 40 scenes were annotated by the two experimenters who were present at the recordings, based on audio and video, for 13 sound events (Table 5.1). The start and stop times of trains and subway trains, and of some speech, singing and screams, were only indicative, because it was hard to denote the exact times when these events became loud enough to be detectable. Furthermore, to assign a single word to a sound event is often not straightforward. For example, to be able interpret a sound event as either speech or a scream, even given a clear scenario, depends on knowledge of the expressing person and the situation, and one word is usually not sufficient to describe an event. However, we chose to annotate the sound events with one word only. As a result, the performance benchmark for the system is one-dimensional but contentious. In section 5.5 we discuss the considerations for annotation procedures and performance benchmarks further.

5.3.2 Setup

The annotations of sound recordings were used to train both a naive Bayes classifier from the Weka toolbox (Witten and Frank, 2005) for the classification of the signal-driven patterns, and the knowledge of the context model. For the naive Bayes classifier, all 40 audio files (each containing one scene) were processed with the signal-driven method described in section 5.2.1. The segregated harmonic complexes with the highest score given by the harmonic complex grouping algorithm—this score is based on the correspondence of the segregated harmonic complex with an ideal harmonic complex, see appendix B—that overlapped with an annotation, were given that annotation as a label. Harmonic complexes without temporal overlap with an annotation were labeled as indefinite. All other harmonic complexes were discarded. Furthermore, pulses and broadband noises that overlapped with an annotation were labeled. From these processed audio files, 40 file pairs were

Table 5.1: Sound events annotated in the audio data recorded at the Amsterdam Amstel station, their occurrences, duration, and duration as part of the total duration of the data set.

Sound event label	Total occurrences	Total duration	Part of total
Speech	521	6 min 10 sec	16%
Scream	290	3 min 33 sec	10%
Singing	82	2 min 32 sec	7%
Subway	40	5 min 22 sec	18%
Kick	26	9 sec	0.4%
Train	15	3 min 20 sec	9%
Subway door signal	14	18 sec	1%
Laughing	12	13 sec	1%
Train whistle	3	5 sec	0.2%
Subway horn	3	3 sec	0.1%
Announcement signal	2	4 sec	0.2%
Birds	2	3 sec	0.1%
Train door signal	1	3 sec	0.1%

generated that contained the feature vectors describing segregated patterns. Each file pair consisted of a file used for training, for which the feature vectors of the segregated patterns from 39 files were used, and a test file, which contained the feature vectors of the segregated patterns 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 (section 5.2.2). On average 18 different types of sequences were encountered in the training set. These sequences are composed of the 13 sound events listed in Table 5.1. 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 was very large, ranging from 2 to 730.

In the test phase, the patterns in the test file are classified with the naive Bayes classifier and used as input for the dynamic network ($\tau = 100$ in equation 4.2, activation threshold $\theta_A = 0.2$, and $K = 0.5$ in equation 5.4). Subsequently, the possible sound events that the pattern can represent are initiated as hypotheses in the net-

work. The weight between the pattern and the event hypotheses is the probability of each event label given by the naive Bayes classifier. If the event cannot be classified and is labeled as indefinite, the weight is set to the default activation of the event. Based on these events, the network forms a hypothesis of a sound sequence, which in turn initiates expectancies of certain sound events that might follow. The results of this integrated approach are the mostly likely events that explain the segregated patterns, given the recognized sound events and their recent history.

The most likely events according to the naive Bayes classifier and the integrated model are compared to the human annotations through 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.[1] Applied to the results of automatic sound recognition in our experimental setup, precision is a measure for the fraction of time the recognitions are correct, and recall is a measure for the fraction of recognitions that are made out of the amount that should have made compared to the annotations.

$$\begin{aligned} \text{precision} &= \frac{TP}{TP+FP} \\ \text{recall} &= \frac{TP}{TP+FN} \\ F &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{5.5}$$

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

5.3.3 Results

The event sequence prediction is based on the classified segregated patterns, and used to select the most likely interpretation for the pattern. Of all 13 types of annotated sound events, 7 are recognized (segregated and labeled) by the Bayes classifier and the integrated model (the segregation algorithm, the Bayes classifier, and

[1] The harmonic mean tends strongly toward the least of the two values. Hence, it penalizes a focus on only one of the two measures.

Table 5.2: *F*-measure, precision, and recall of random classification (R), a Bayes classifier (C), and the integrated model (I) per sound event type.

	F-measure			Precision			Recall		
Sound event label	R	C	I	R	C	I	R	C	I
Singing	0.07	0.28	0.48	0.07	0.26	0.36	0.07	0.31	0.72
Speech	0.16	0.18	0.38	0.16	0.50	0.44	0.15	0.11	0.33
Train	0.09	0.40	0.43	0.09	0.32	0.35	0.09	0.54	0.55
Subway door signal	0.01	0.28	0.26	0.01	0.25	0.21	0.01	0.33	0.33
Subway	0.17	0.52	0.57	0.18	0.54	0.62	0.17	0.49	0.53
Kick	0	0.28	0.24	0	0.26	0.65	0	0.30	0.14
Scream	0.09	0.36	0.44	0.10	0.36	0.36	0.09	0.36	0.56

the context model). These 7 types of sound events were most frequently annotated (see Table 5.1). Hence, the Bayes classifier and the context model can learn them more reliably than events that occur infrequently. Table 5.2 shows the *F*-measure, precision and recall of the recognitions made by the Bayes classifier (C) and by the integrated model (I) for the 7 sound event types. These measures can be compared to a random classification (R), which is based on the amount of time that the sound events are annotated, that is, the a priori probability that the sound events are encountered. Figure 5.3 displays the overall results of both models.

On average, the context model improves the *F*-measure compared to the signal-driven classification, mostly through an increased recall of the sound events that have more harmonic content (singing, screams and speech), because these types of events are more likely to be of the same type as their surrounding events. For example, 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 event and the precision of the speech event increase. However, the increase in precision of harmonic sound events is moderated by some erroneous changes in other sound events. As a result, the overall precision does not increase as much as the recall.

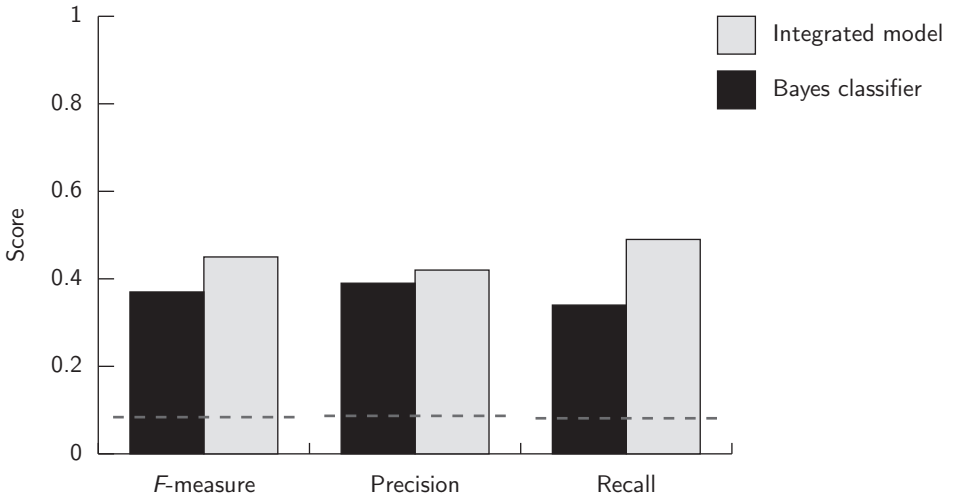


Figure 5.3: Overall results of the Bayes classifier and the context model integrated with the Bayes classifier on the data set recorded at the Amsterdam Amstel train station. The dashed lines show the overall performance of random classification, which is based on the average amount of time that the sound events are annotated.

5.4 EXPERIMENT 2

This experiment demonstrates that the proposed methods can be used to recognize sound events in a sonic environment given a predicted location. Furthermore, both the environment and the recorded events are uncontrolled in the data set of this experiment. As a consequence, the diversity of the recorded events is enhanced compared to the first experiment. In the next section we describe the data set that is used in the experiment. Subsequently, the setup of the experiment is explained, and in the last part we present the results.

5.4.1 Data

The data 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 three minute recordings at six different locations: a railway station platform, a pedestrian crossing with traffic

Table 5.3: Examples of sound events annotated in the data set recorded in Assen, their occurrences, duration, and duration as part of the total duration of the data set.

Sound event label	Total occurrences	Total duration	Part of total
Bird	238	17 min 6 sec	11%
Bike	30	2 min 22 sec	2%
Rooster	16	43 sec	0.5%
Horn	8	11 sec	0.1%
Shopping bag	1	7 sec	<0.1%

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.

All the recordings were annotated by two students separately. The average agreement between the two annotators of one group was determined with the F-measure (equation 5.5): $\overline{F} = 0.46$ with a standard deviation of 0.25. These two annotations were merged, such that equal labels did not overlap, but became one instance (a union in set theory). We examined the resulting merged annotations, and adjusted them when necessary. However, we did not introduce new annotations.[1] We ensured that the names of events were uniform across all the files to prevent the context 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 sound events. However, half of these sound events were annotated less than 5 times, while just a few events comprised most of the annotations. A few examples of annotated events are given in Table 5.3, ranked according to their frequency in the complete data set.

5.4.2 Setup

The experiment was designed in the same manner as the first experiment (section 5.3.2). We used a nearest-neighbor classifier instead of the naive Bayes classifier

[1] An exception was made for the annotations of one group, which we had to complete because they were sloppy or omitted.

to label the segregated patterns. For the nearest-neighbor classifier, all 44 audio files were processed with the signal-driven method described in section 5.2.1. The segregated harmonic complexes with the highest score given by the harmonic complex grouping algorithm—this score is based on the correspondence of the segregated harmonic complex with an ideal harmonic complex—that overlapped with an annotation were given that annotation as a label. Harmonic complexes without temporal overlap with an annotation were labeled as indefinite. All other harmonic complexes were discarded. Furthermore, pulses and broadband noises that overlapped with an annotation were labeled. From these processed audio files, 44 descriptive file pairs were generated that contained the feature vectors describing segregated patterns. Each file pair consisted of a file used for training, for which the feature vectors of the segregated patterns from 43 files were used, and a test file, which contained the feature vectors of the segregated patterns 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 context model (section 5.2.2). The information used to train the weights between the locations and the sound events is summarized in Table 5.4. Furthermore, it shows a few examples of the connection strength between the locations and some sound events. On average 21 sound events were encountered at each recording location in the training set. Furthermore, an average duration of 80 seconds per sound event at a location was used to train the weights in the knowledge network. The spread of duration per sound event was very large, as can be seen in Table 5.3.

In the test phase, the segregated patterns in the test file, which are classified by the nearest-neighbor classifier, are used as input for the dynamic network ($\tau = 100$ in equation 4.2 and the activation threshold $\theta_A = 0.2$). Subsequently, the possible sound events that the pattern can represent according to the learned knowledge are initiated as hypotheses in the network. The weights between the pattern and the event hypotheses are the probabilities of each event given by the nearest-neighbor classifier. If the event cannot be classified and is labeled as indefinite, the weight is set to the default activation of the event. Based on these event hypotheses, the network forms a hypothesis of the location, which in turn initiates expectancies of certain sound events that might follow. The results of this integrated approach are the mostly likely sound events that explain the segregated patterns, given the recog-

Table 5.4: Locations. $\overline{N}$ is the average number of sound events annotated at a location, $\overline{T}$ is the average of the total duration per annotated sound event at a location in seconds, and $\overline{s_T}$ is the average spread of the event durations at a location. The two weights $w_{l,\text{train}}$ and $w_{l,\text{bus}}$ are examples of learned connection strengths between locations (l) and sound events.

Location	$\overline{N}$	$\overline{T}(s)$	$\overline{s_T}$	$w_{l,\text{train}}$	$w_{l,\text{bus}}$
City center	23	64	152	0	0
Graveyard	15	100	157	0	0
Museum	24	68	93	0	0.02
Traffic lights	17	87	181	0	0.15
Train station	26	79	111	1	0.02
Walking	21	81	170	0	0.04

nized sound events and their predicted location. The most likely events according to both the nearest-neighbor classifier and the integrated model are compared to the annotations through the F-measure (equation 5.5).

5.4.3 Results

The success of the context model as it is applied in this study is dependent on whether the recording location prediction is correct. The results of the location predictions of the test files are listed in Table 5.5. The number of test files at each location is in parentheses behind the location name. The location predictions of the 7 test files of recordings during walking are not included, because they cannot be assigned to a single location. The top 1 indicates the amount of time the location predictions are correct on average for a specific location. The model has 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 the some sound events that are specific for those locations, such as train sounds for the train station, are segregated and correctly classified. In contrast, many of the other sounds that can be segregated and classified by the nearest-neighbor algorithm, such as cars and speech, are generic, and can be heard at any of

Table 5.5: Results of location predictions: the average amount of time that the location prediction was correct per location. The number of files recorded at each location is shown in parentheses.

Location	Top 1	Top 2	Top 3
City center (7)	1%	2%	2%
Graveyard (7)	1%	1%	15%
Museum (8)	6%	87%	93%
Traffic lights (7)	2%	8%	59%
Train station (8)	90%	98%	98%

the locations. Therefore, the location prediction is not reliable in many test files.

The location prediction is based on the classified segregated patterns, and used to select the most likely label for the pattern. Of all 54 annotated sound events, 11 events are recognized (segregated and labeled) by the nearest-neighbor classifier and the integrated model (the segregation algorithm, the nearest-neighbor classifier, and the context model). These 11 sound events are most frequently annotated. Hence, the nearest-neighbor classifier and the context model can learn them more reliably than sound events that occur infrequently. Table 5.6 shows the *F*-measure, precision and recall of the identifications made by the nearest-neighbor classifier (C) and by the integrated model (I) for the 11 sound events. These measures can be compared to a random classification (R), which is based on the amount of time that the sound events are annotated, that is, the a priori probability that the sound events are encountered. Therefore, the results of the random classification are relatively high on sound events that are annotated often and long, such as speech and footsteps.

On average the context model slightly improves the *F*-measure compared to the signal-driven classification, mostly through an increased recall, which means that more correct instances of annotations are found than with the nearest-neighbor classifier (Figure 5.4). For both models the performance on this data set is lower than the performance on the data set recorded at the Amsterdam Amstel train station, because this data set is more divers, and recorded in less controlled conditions. Furthermore, the first data set is annotated by two people, compared to 14 in this data set, resulting in more diverse annotations.

Table 5.6: *F*-measure, precision, and recall of random classification (R), a nearest-neighbor classifier (C), and the integrated model (I) per sound event.

	F-measure			Precision			Recall		
Sound event label	R	C	I	R	C	I	R	C	I
Bird	0.109	0.005	0.053	0.112	0.093	0.198	0.106	0.003	0.031
Braking train	0.018	0.193	0.194	0.019	0.274	0.188	0.018	0.149	0.201
Bus	0.026	0.053	0.055	0.026	0.201	0.029	0.026	0.031	0.423
Car	0.195	0.499	0.408	0.203	0.594	0.427	0.184	0.431	0.391
Footsteps	0.173	0.008	0.104	0.181	0.091	0.358	0.166	0.004	0.061
Passing train	0.005	0.570	0.612	0.005	0.416	0.463	0.005	0.906	0.906
Rain drops	0.044	0	0.052	0.045	0	0.135	0.044	0	0.033
Speech	0.127	0.023	0.111	0.131	0.223	0.124	0.123	0.012	0.100
Starting train	0.019	0.022	0	0.019	0.018	0	0.019	0.028	0
Truck	0.057	0.021	0.013	0.057	0.045	0.010	0.056	0.014	0.017
Wind	0.129	0.071	0.138	0.134	0.134	0.135	0.125	0.048	0.141

5.5 CONCLUSION

In the previous sections we have demonstrated that a model that combines both signal-driven algorithms and contextual knowledge in the form of predicted recording locations or event sequences, improves the recognition of sound events in a real-world environment compared to an exclusively signal-driven method. Especially the events that have similar audio patterns, and hence rely more on context for their interpretation, for example screams, speech and singing in the first experiment, are recognized better in the integrated approach. Sound events that are already recognized well by the signal-driven algorithm, such as trains and cars, gain little improvement from the context model. Finally, sound events that occur infrequently, and hence have few training examples, and events that are not yet captured well by the audio features vectors, show a small performance reduction.

The overall results (especially of the second experiment) may not seem impressive, but this is partly explained by the performance measure. The *F*-measure is based on the temporal overlap of the annotations and the labeled patterns. Therefore, it is dependent on both the annotations and the segregation algorithm. Annotating sound is a complex process, which is demonstrated by the low inter-

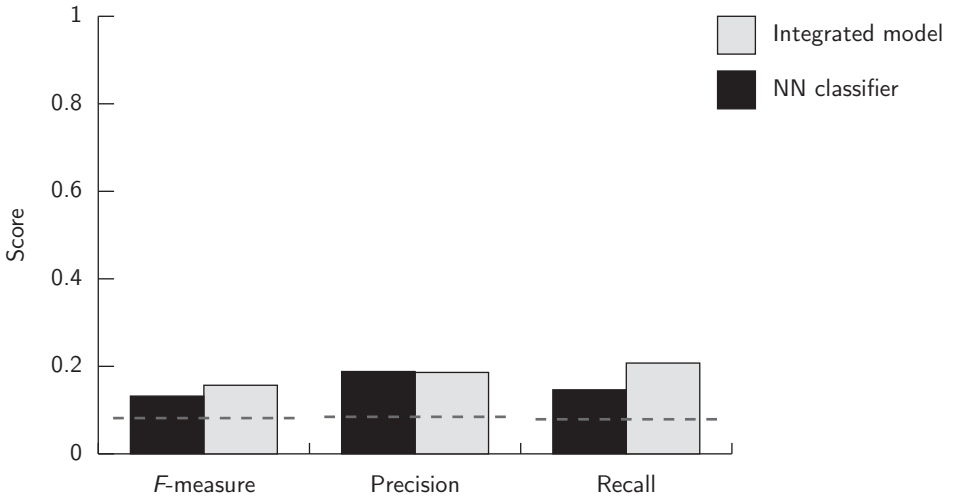


Figure 5.4: Overall results of the nearest-neighbor (NN) classifier and the context model integrated with the nearest-neighbor classifier on the data set recorded at different locations in Assen. The lower dashed lines show the overall performance of random classification, which is based on the average amount of time that the sound events are annotated.

annotator agreement ($\overline{F} = 0.46$). The annotators did not only use information in the sound, 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 sound or on their knowledge. Some annotated sound events can even hardly be recognized by a human who has to rely on the audio signal alone.

In contrast, the segregation algorithm relies solely on the audio signal. This signal is uncontrolled and thus very challenging for the algorithm that needs to segregates relevant parts. For example, people annotate sound events that occur in a sequence, such as footsteps and birds, continuously even though they are interrupted.[1] In contrast, the segregation algorithm has to segregate every single occurrence before it can be recognized. As a consequence, the precision of both models on these types of sound events can never be high. Furthermore, especially the recordings in the second experiment contain a wide variety of sounds events,

[1] The random classification outperforms the models on a few of these sound events, because it is based on a priori probability of occurrence determined from the annotations.

most of which occur only a few times in all the recordings.[1] To be able to learn the patterns of a sound event, a machine learning algorithm (like Bayes or nearest-neighbor) needs more examples than were available of most sound events in the data set in these experiments.

These observations demonstrate that modeling context information is essential to achieve robust event recognition in real-world environments. Indeed we have shown that context, in the form of recent history of sound events, improves sound event recognition, even though it is so far only based on knowledge derived from the human annotations of the audio signal. Additionally, the recording location is not predictive for many generic sound events, such as speech and cars[2], which occur most often, and are classified best by the nearest-neighbor classifier. In other words, the infrequent events are the events that are good predictors of a location, while these are the hardest events to learn, because they are infrequent. However, the context model is not limited to process acoustic information. In the next chapter we show that the context model can also be used to process ambiguous visual information. 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 integrated model provides a new strategy to analyze sonic environments by recognizing sound events. Because these sound events are also based on knowledge of the context, they are a first approximation of meaningful auditory events. However, to improve the recognition of these events in the complexity of real environments, we require a combined development of segregation algorithms and models that can include non-acoustic factors. Furthermore, we will study human perception in parallel, so we can validate the model for soundscape analysis. Conversely, the development of a system to analyze a soundscape automatically might increase our understanding of human soundscape evaluation.

---

[1] The average amount of time that each sound event is annotated is 2%. Excluding the few sound events that occur most often (birds, cars, footsteps, speech, and wind), this amount is less than 1%.

[2] In the knowledge network of the second experiment, the weights between speech and cars and all locations were 0.

6

Automatic analysis of ambiguous visual information

The content of this chapter has been published as Niessen, M. E., Kootstra, G., De Jong, S., & Andringa, T. C. (2009). Expectancy-based robot navigation through context evaluation. In Proceedings of the 2009 Interenational Conference on Artificial Intelligence (pp. 371–377).

Artificial agents that operate in a real-world environment have to process an abundance of information, which may be ambiguous or noisy. We present a method grounded in cognitive research that keeps track of sensory information, and interprets it with knowledge of the context. We test this model on visual information from the real-world environment of a mobile robot in order to improve its self-localization. We use a topological map to represent the environment, which is an abstract representation of distinct places and the connections between them. Expectancies of the place of the robot on the map are combined with evidence from observations to reach the best prediction of the next place of the robot. These expectancies make a place prediction more robust to ambiguous and noisy observations. Results of the model operating on data gathered by a mobile robot confirm that context evaluation improves localization compared to a signal-driven model.

6.1 INTRODUCTION

Artificial agents that operate in a real-world environment are confronted with additional challenges compared to agents that operate in a controlled or simulated environment. They have to process an abundance of information, of which not everything is necessarily relevant for their goal. Moreover, sensory information may be ambiguous or noisy. To make sense of its environment, an agent needs to identify and structure the sensory information it gathers. We developed a method grounded in cognitive research that keeps track of sensory information, and interprets it with knowledge of the context (chapter 4).

Applications of cognitive research, such as handwriting recognition (Côté *et al.*, 1998) and information retrieval (Crestani, 1997; Van Maanen *et al.*, 2010), often employ a spreading activation semantic network to recognize a particular item or retrieve specific information. Spreading activation networks are based on a model of human memory (Quillian, 1968). They are realized as connected nodes that represent pieces of information or concepts, and the vertices represent the prior probabilities that the nodes are encountered together. Spreading activation networks are typically static, because the data in these application domains can be accessed completely and simultaneously. In contrast, for agents operating in a dynamic environment the available information continuously changes.

To deal with continuous data, we apply a context model that manages a dynamic network. This dynamic network is similar to a spreading activation network, but instead of being static, it is updated when new data are encountered. The model continuously updates its current state, based on sensory input and knowledge of the context. The context model has been applied to recognize sound events (see chapter 5), but is designed to manage any type of sensory input. We will show in this chapter that it can also be applied to visual information from the real-world environment of a mobile robot.

A basic task for an autonomous mobile robot is to build a map of its environment for self-localization. For this reason, simultaneous localization and mapping (SLAM) has received considerable attention in the last decade. Most SLAM approaches use range or vision sensors to construct a detailed metric map of the environment (Thrun *et al.*, 2005). These maps contain the Cartesian coordinates of many structural features present in the environment. Other approaches build

topological maps of the environment (Vasudevan *et al.*, 2007). Instead of representing the environment in detail, it is represented more abstractly in topological maps, as distinct places and the connections between them. The advantage of such an abstract representation is that it is less susceptible to noise, and ambiguous observations and situations. Moreover, it results in a computationally less demanding system.

In topological mapping, a rough estimate of the location of the robot can help to form an expectancy of the path of the robot. This expectancy can be combined with evidence from observations to form a hypothesis of the place of the robot. Furthermore, an expectancy of the place of the robot can resolve ambiguous observations. In this way, the place in a topological map where an observation is made can be considered as the topological context of that observation. When the robot is moving and making observations, an evaluation of the context can improve its localization. The evaluation of the context entails that the recent history of visited places is used to predict the place that follows. Furthermore, using knowledge of the topological context makes localization more robust to noise in the observations.

In the next section we describe the design of the model, and how it processes observations made by a mobile robot. In section 6.4 we present the results of two experiments that are described in section 6.3. The first experiment demonstrates that the model is more robust to noise when the topological context is used. The second experiment shows that predictions in real data with many ambiguous observations and noise are also better with context evaluation than without. We end with a discussion on the performance of the model and give an outlook on future work.

6.2 METHODS

The model we present processes visual input of a moving robot. These visual observations, which are explained in section 6.2.1, provide evidence about the place of the robot. However, ambiguous or noisy observations can lead to erroneous place predictions. To improve these predictions, contextual information about the environment is learned in a supervised training phase and stored in a static knowledge network. In the operation phase this knowledge is used in a dynamic network,

which computes expectancies of the place of the robot.

The knowledge about the environment, in the form of nodes in the knowledge network and the strength (weight w) of the connections between them, is computed in the training phase. We refer to this knowledge as long-term memory, since it reflects stable knowledge. Therefore, it is stored as a static network, which is constructed from learning relations in the training data. This knowledge network is similar to semantic networks used in information retrieval. In section 6.2.2 we describe in more detail how the knowledge network is created.

In contrast to the knowledge network, the dynamic network reflects short-term memory. Information represented by nodes in this network is added and forgotten more quickly, since the nodes pertain only to the current state of the robot. Nodes in the dynamic network are called hypotheses, because they represent possible explanations for input data. The dynamic network has three levels that all represent a different type of information: hypotheses of observations, landmarks, and places in the environment. Figure 6.1 shows an example of a dynamic network at one moment, namely when observation 2 has been made. The network configuration represents the knowledge of the environment at that moment. This knowledge consists of two observations, their connections to landmarks hypotheses, and the connections of the landmarks to hypotheses of places in the environment. In section 6.2.3 we explain the construction of the dynamic network, and how topological context is used to compute expectancies of the place of the robot.

6.2.1 Observations

The robot (a Pioneer 2 DX mobile) uses a video camera to navigate in its environment. Visual interest points are detected in the camera images, which serve as landmarks to represent the environment. The interest points are detected and described using the scale-invariant feature transform (SIFT, Lowe, 2004). The SIFT algorithm detects points that stand out from their surroundings. These points are described using histograms of gradients. A drawback of SIFT is that it results in a large number of interest points, many of which are not re-detected in subsequent images. Therefore, we use a visual buffer to test the stability of the interest points over a number of successive images (Kootstra *et al.*, 2009). Only interest points that are stable enough are used as landmark observations. The descriptor of an

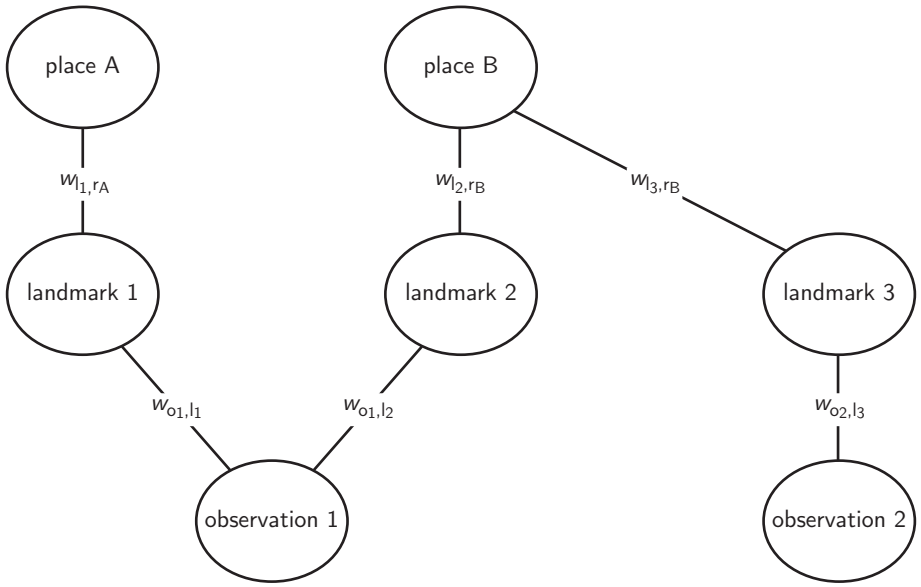


Figure 6.1: Example network configuration at one instant, of two observations that are matched to three landmarks, each in turn connected to a place.

observation is then compared to that of previously observed landmarks. Based on the descriptor distances, the observation is matched with one or more landmarks or labeled as a new landmark.

The data set used in one of the two experiments (section 6.3.2) was collected by the robot while it drove a closed loop of eight by ten meters in an office-like environment. The data was logged by the robot while driving four laps. The map of the loop was manually divided into nine places, as depicted in Figure 6.2. Half of the data set, that is, the observations made in the first two laps, is used to determine which landmarks are observed in which place. The other half is used to test the model. Because of the variability of the images in different laps, the robot might have observed landmarks in the last two laps that are not present in the training data.

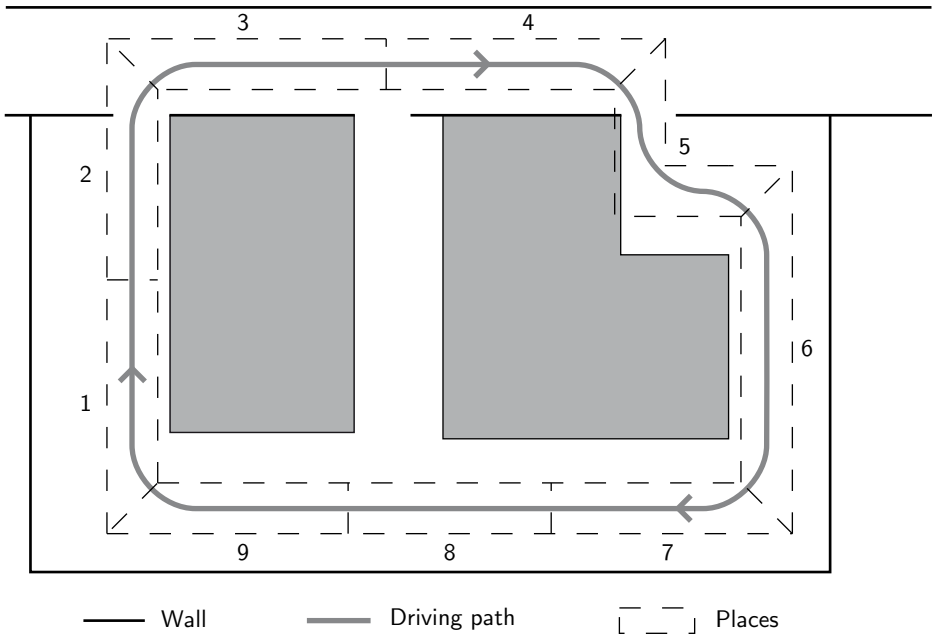


Figure 6.2: Environment where the robot drove four laps. The size of the loop is 8 by 10 meters, divided into nine places. The gray area consists of objects the robot cannot drive through.

6.2.2 Knowledge network

Three classes of information are stored in a knowledge network: the descriptors of the landmarks, the relations between the landmarks and the places in the environment, and the transitions between the places. This knowledge network represents the context, which is slowly changing or invariant. Therefore, it is referred to as the long-term memory of the model.

The connection strengths between landmarks and places in the training data are 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 (term frequency) with its general frequency in other documents (inverse document frequency). Hence, the term is important for a document if it occurs often in that document and infrequently in other documents. Since the connection

strength (weight w) between a landmark and a place should reflect the specificity of the landmark to that place, we adopt the term-weighting approach. The landmarks can be treated as terms, and the places as documents. Accordingly, the weight of the connection between landmark l and place r is:

$$w_{r,l} = w_{l,r} = \mathrm{tf} \cdot \frac{\log_{10} N - \log_{10} n}{\log_{10} N} \tag{6.1}$$

where N is the total number of places, n is the number of places in which landmark l is observed, and the normalized term frequency is given by:

$$\mathrm{tf} = \frac{f_{l,r}}{\sqrt{f_l}}, \tag{6.2}$$

where $f_{l,r}$ is the observation frequency of l in r, and f_l is the total observation frequency of l.

The connections between observations and landmarks are not stored in the knowledge network, because all observations are unique. Therefore, the weights of these connections are computed at the moment when an observation is made, both in the training and the operation phase.[1] The connection strength between an observation and a landmark should represent the likelihood of a correct matching between their descriptors. If these descriptors are far apart, the observation and landmark are less likely to have been matched correctly. Therefore, the weight of a connection between an observation and a landmark is inversely related to the distance between their descriptors:

$$w_{l,o} = w_{o,l} = 1 - \frac{d}{\theta_d}, \tag{6.3}$$

where d is the distance between the descriptor of observation o and landmark l, and θ_d is the maximum distance at which an observation is still matched to a known landmark.

The transition probability that the robot moves from one place to another is calculated by normalizing the number of times the robot moves from one place

[1] Observed information is not necessarily always unique. In other domains or applications areas it could be useful to store observations in the knowledge network. However, in the presented application it would be useless to do so.

to another in the training data (the first two laps). As can be seen in Figure 6.2, the robot can move within place i or move from place i to place $i \pm 1$. Since the robot is driving the loop in one direction, the transition probabilities to all places other than i and $i+1$ are generally zero. However, there are a few exceptions when no observations are made in a place in one of the laps, and thus the probability to move to $i+2$ is greater than zero. The complete matrix of probabilities serves as the topological context that helps to compute an expectancy about the next location of the robot.

To summarize, the knowledge network consists of the matrix with the a priori transition probabilities between all places. Furthermore, it stores the labels of all landmarks that are observed in the training data, along with their connections to the places in which they are observed.

6.2.3 Dynamic network of hypotheses

Once the knowledge network is fully trained after the learning phase, it is used in the operation phase, together with evidence from observations, to predict the place of the robot. The algorithm for the construction and updating of a dynamic network is summarized in Table 6.1 (see chapter 4 for a detailed description). Every level in the network consists of hypotheses of a single type of representation (see Figure 6.1). The landmark observations are the lowest level of the dynamic network. As described in section 6.2.1, observations are matched to one or more previously observed landmarks, or labeled as a new landmark, which are at the middle level. The highest level in the network holds hypotheses of places in the environment.

Each node in the network represents a hypothesis of one of the three different types of representation. When an observation is made, a hypothesis is added to the dynamic network (step 1). Next, its matched landmarks (that are stored in the knowledge network[1]) are initiated as hypotheses, and they are connected to the observation hypothesis (step 2). Subsequently, these landmark hypotheses retrieve their place connections from the knowledge network. These places are also initiated as hypotheses (step 3) and connected to the landmark hypotheses that initiated

[1] The current version of the model only processes known landmarks in the operation phase. The possibility to add new landmarks will be discussed in section 6.5.

Table 6.1: Algorithm for updating the dynamic network configuration at times when observations are made by the robot.

For all observations at time t:	
1.	Add observations $O(t) = \{o_{t,i}\}$ to the network
2.	For each $o_{t,i} \in O(t)$ add matched landmark hypotheses $\{l_{t,j}\} \in L(t)$ and connect them with strength $w_{i,j}$
3.	For each new landmark $l_{t,j} \in L(t)$ add appropriate places r_k not yet present in the set of active places R
4.	Connect each new landmark $l_{t,j}$ with the appropriate place r_k with strength $w_{j,k}$
5.	Spread signal-driven activation
6.	Spread context-based activation
7.	Evaluate activation values

them (step 4). Every time new observations are made, the network is updated and the dynamics change.

Activation spreading

After the connections in the network are updated, the activation of the observation hypothesis spreads through the network (see section 4.3). The input activation first spreads upward to the place hypotheses at the highest level in the network, and is called signal-driven spreading (step 5). Subsequently, the activations of the place hypotheses spread downward to other connected hypotheses, for example landmarks in the same place that are observed previously. We call this context-based spreading (step 6). As a consequence of context-based spreading, a landmark hypothesis of a particular observation can be reinforced by later observations. For example, in Figure 6.1 the first observation is matched to landmarks 1 and 2, where landmark 1 lies in place A and landmark 2 in place B. Another landmark observation made in place B will increase the support for the hypothesis that the first observation was of landmark 2, and not of landmark 1.

Activation evaluation

After the activation has spread through the network, the activation value of each hypothesis is evaluated (step 7). The activation evaluation is different for different

types of hypotheses. The activations of the hypotheses that are not at the highest level in the network are normalized (equation 4.3). However, the place hypotheses have an expected activation value, like sound events in fixed sequences (section 5.2.2), because the order in which the robot drives through the environment is not random. Therefore, the activations of the place hypotheses at the highest level are a weighting of evidence from the input and an expected value.

The expected activation of place hypotheses represents the expectancy to be at a place given the context. It is calculated using the information about the place transitions in the environment (Figure 6.2). The expected activation of place i is the sum of all possible options to drive to place i:

$$\hat{A}_i(t) = \sum_j f_j(\Delta t) A_j(t - \Delta t) P(j \rightarrow i) P(j) \text{ for } i, j \in R, \tag{6.4}$$

where $A_j(t - \Delta t)$ is the previous activation of place hypothesis j, multiplied with a decay $f_j(\Delta t)$, $P(j \rightarrow i)$ is the transition probability from place j to place i, including $j = i$, the probability to stay in the same place. Finally, $P(j)$ is the a priori probability to be in place j, and R is the subset of hypotheses that represent places.

The a priori transition probabilities from equation 6.4 are retrieved from the knowledge network. The probabilities are adjusted in the dynamic network of hypotheses, because the probability that the robot leaves a place increases as it is longer in that place. More specifically, the probability of staying in the same place decreases as a function of the age T_i (how long it is active) of the place hypothesis: $P(i \rightarrow i)(T_i) = P(i \rightarrow i)^{T_i}$. The probabilities to move to other places are increased proportionally to their a priori connection strength. For example, suppose the initial transition probability between place A and place B is 0.2, and the probability of staying in place A is 0.8. After the robot has observed landmarks in place A at four subsequent times, $P(A \rightarrow A) = 0.8^4 = 0.4$ and $P(A \rightarrow B) = 0.6$. When the robot returns to the same place, the probabilities are re-initialized to the probabilities in the knowledge network.

The expected activation is combined with evidence from the current input to compute the activation evaluation of the place hypotheses:

$$A_i(t) = \hat{A}_i(t) + K \left(\frac{n_i(t)}{\max n(t)} - \hat{A}_i(t) \right) \text{ if } i \in R, \tag{6.5}$$

where $\hat{A}_i(t)$ is the expected activation according to equation 6.4, $n_i(t)$ is the input activation of i as calculated in equation 4.1, $n(t)$ is a list with the input activations of all active place hypotheses, and K is the gain factor. The gain factor is dependent on the noise in the observations. If the observations are very reliable, its value should be high. However, the current data set is relatively noisy. Therefore, the gain factor is set to 0.25, which entails that the model responds relatively slowly to new observations, and is guided more by expectancies.

The final activation values of all active place hypotheses are compared, and the one with the highest activation is the current best hypothesis of the place of the robot. Hence, the sequence of best hypotheses at each update gives the estimation of the model of the path of the robot.

6.3 EXPERIMENTS

To illustrate the benefit of context evaluation in robot localization, we show the place predictions of two models. In the first model the predictions are based on instant observations alone, which implies that only information from the knowledge network is used. Accordingly, context-based spreading is not applied, because the signal-driven model does not remember previous predictions. In other words, hypotheses of the place of the robot are deactivated after the signal-driven activation spreading. In the second model, the context-based model, the place prediction is based on a combination of instant observations and expectancies, which are computed through context evaluation, as discussed in section 6.2.3 ($\tau = 67$ in equation 4.2, activation threshold $\theta_A = 0.05$, and $K = 0.25$ in equation 6.5).

We discuss the results of both models running on two types of data. In the section 6.3.1 we present an experiment with simulated data, which can be controlled in their complexity. The simulated data are a simplification of the real data described in section 6.2.1. The experiment with the real data is discussed in section 6.3.2.

6.3.1 Simulated data

We generated a data set to measure the performance of the model on data with different levels of noise. The noise simulates observations that are so similar that they are matched to the same landmark, although the observations are made at

distinct places. These types of ambiguous observations occur often in the real data due to reoccurring objects and structures in office-like environments. At every time step one observation is simulated, which is matched to one landmark. The distance between the descriptor of the observation and the landmark is set to the same value for all observations. In the first lap all 240 landmarks are observed and connected uniformly to one of eight places. No noise was applied in the training part of the data, the first two laps, so there are no ambiguous landmarks in the a priori knowledge network. In the test data we applied a varying amount of noise on the landmarks. When no noise is applied to the data, the test set is identical to the training set. As the amount of noise increases, the place at which a landmark is observed becomes more random, until it is completely random at a noise level of 100%.

6.3.2 Real data

In the real data, as described in section 6.2.1, 225 unique landmarks are observed in the first two laps (the training data). In the operation phase, 107 of these landmarks are re-observed and used as input to the dynamic network. In the operation phase 114 new landmarks are detected, which are not processed by the current version of the model. Of the landmarks in the knowledge network 24% is ambiguous, that is, these landmarks are observed in more than one region in the training phase. The real data are quite challenging, because they contain noisy and erroneous observations, hold many ambiguous landmarks, and landmarks that are unequally distributed in the environment.

6.4 RESULTS

The results of the model on the simulated data are shown in Figure 6.3. Since the model keeps track of all hypotheses, there is a list of hypotheses with a decreasing activation value, not only a single winner. Hence, it is possible that the true place is not the best hypothesis, but the second best. Therefore, the performance of the model can be evaluated not only by comparing the true place to the best place hypothesis, but also to the top two or top three. Figure 6.3 depicts the best result (top one) for the context-based model and the signal-driven model, and the top two and three of the context-based model. The results of the signal-driven model are iden-

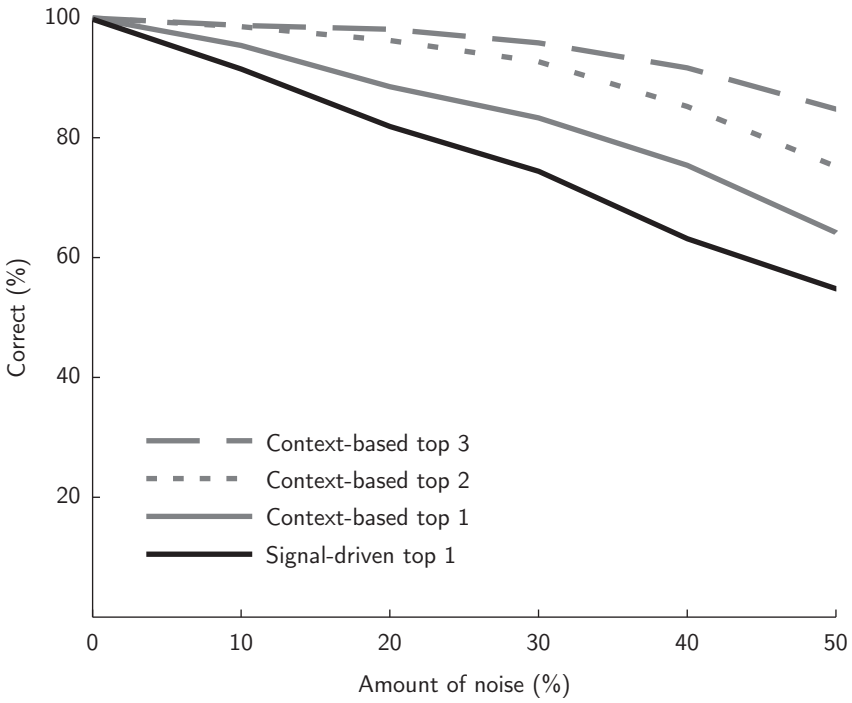


Figure 6.3: Results of the model tested on data with a varying amount of noise. The single best result of the context-based model and the signal-driven model are shown, and the top two and top three of the context-based model.

tical for the top one, the top two, and the top three, because the simulated data set contains only one observation per time step, resulting in one possible hypothesis.

As expected, the signal-driven model performs at chance level. When an incorrect observation is made, the place prediction is also false. The context-based model performs better than the signal-driven model, especially for low amounts of noise ($< 50\%$).[1] In the experiment on the real data, of which the results are depicted in Figure 6.4, the context-based model also out-performs the signal-driven model. The difference between the score of the best hypothesis of both models is not very large, but consistent in multiple tests. However, the high scores on the

[1] High levels of noise are not included in the figure, because the results are less meaningful if the noise is more prominent than the observations.

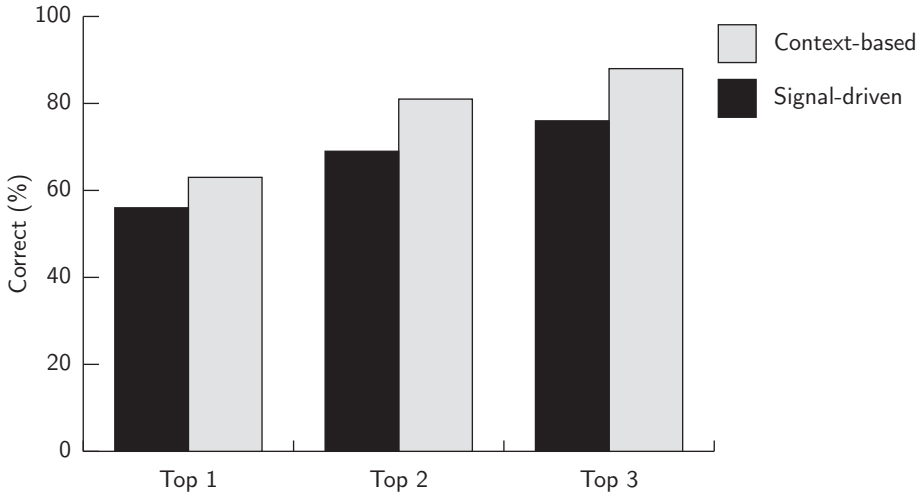


Figure 6.4: Results of the signal-driven and the context-based model on data collected by a moving robot.

top two and three are promising for future improvement.

It should be noted that the predictions of both models are based solely on visual observations, and odometric information is ignored. Therefore, the results of the two models can be compared by their performance on visual information, and we can show the advantage of the context-based model. If one would aim at a best possible robot localization, odometric information should be included.

6.5 CONCLUSION

We presented a model that dynamically manages a spreading activation network. This network represents the environment of an agent based on sensory information and knowledge of the environment. To test the applicability of the model in a real-world environment, we tested it on visual observations gathered by a mobile robot, with the goal to improve its localization. Learned knowledge about the environment of the robot is used to compute expectancies of its location. These expectancies are combined with instant observations to form a prediction of its location. Including expectancy in the prediction enhances the stability of the model, since it prevents unexpected landmarks from disrupting the place prediction.

The information about the environment is learned in a supervised training phase, and stored in a knowledge network, the long-term memory of the model. The short-term memory is represented by a dynamic network. Hypotheses in the dynamic network are more transient, because they represent the current state of the robot. The network and the deduced location prediction are updated when the robot gathers new evidence about its environment. The results of the experiments confirm that context evaluation improves the performance compared to signal-driven evaluation on both the simulated and the real data.

Although the context-based model outperforms the signal-driven model, the difference on the top one in the experiment on the real data is not very large. This can be explained by the fact that more than half of the landmarks that the robot encounters during the operation phase are new. Hence, the information on which the model can base its prediction is limited. Therefore, it will be useful to integrate an algorithm in the model that includes new landmarks in the knowledge network during the operation phase. For example, the growing-when-required (GWR) network of Marsland *et al.* (2002) adds new nodes to a network based on the (mis-) match between the data and the network. Such an algorithm would make it possible to learn new information during the operation phase. Furthermore, incremental learning can be used to update existing connections based on new observations.

Another possible improvement can be made in the determination of expectancies. In the current version of the model we only update the network when observations are made. This can pose problems to the model, especially when the data is not equally distributed over the environment, causing some places to be poorly represented by landmarks. Based on temporal and odometric information, expectancies of the path of the robot can be made even without observations.

In conclusion, the presented model can improve robot localization through context evaluation. It is computationally efficient and needs little memory storage. Therefore, it can be easily scaled to larger environments. Moreover, the model is general, because the sensory information in the model is not limited to visual observations. Hence, it can be used for state estimation in other domains (see chapter 5), or even combine information from different modalities to make predictions.

7

General discussion

7.1 CHALLENGES

Environmental sound event recognition is a young research area compared to sound recognition aimed at a specific type of sound, in particular speech. As a consequence, methods to deal with the complexities of real-world environments are still being developed, and not yet as accomplished as methods used in single-type sound recognition. The methods that have proven to be successful in single-type sound recognition have been exported to the field of environmental sound event recognition. However, the two problems are qualitatively different. More specifically, the context in single-type sound recognition is provided by the problem definition, and the type of sound to be recognized is a priori defined. For example, contextual information in the form of grammar rules can be applied in automatic speech recognition to improve recognition of ambiguous signal information. Moreover, methods to deal with a distorted signal caused by transmission effects can rely on the presence of speech. In contrast, both the context and the type of sound in environmental sound event recognition are variable.

To advance in environmental sound event recognition the variability of the sound events and the environment needs to be accounted for. The methods used in single-type sound recognition rely on stable and known acoustic properties of the audio signal. Therefore, they are not flexible enough to meet the requirements of real-world environments. Instead, we need methods from the field of computational auditory scene analysis that segregate components that are likely to constitute a single event. These segregated components provide hypotheses about sound events that have to be interpreted with knowledge of the environment. The semantics of diverse environments and sound events have to be learned, so that a system for environmental sound event recognition can work in variable environmental contexts. For example, although knowledge about the context in the form of a statistical language model is used in automatic speech recognition, it relies on the temporal structure of the input signal. Therefore, it cannot generalize to other domains with a different structure. In contrast, the context model presented in this thesis is more flexible, because it can learn different types of structure in the environment.

Our method to integrate the semantics of the sound events and the environmental context in automatic sound event recognition is based on a model of human

memory. The learned knowledge that is stored in a network represents long-term memory, because it is assumed to be stable. The semantics (in the form of linguistic labels) in the model are learned from human annotations. Therefore, the representation of the sonic environment is independent of audio descriptors. As a consequence, the model can be applied to other types of information as well, such as visual information or positioning information. This generic representation of the environment is more robust to changing conditions, such as transmission effects, than a representation that is based on the acoustics of the sound events. An acoustic representation relies on the quality of the signal processing techniques to select information in the signal that is specific for the sound event. However, this is difficult in a real-world environment where events can be (partly) masked by other sound events and distorted by transmission effects. Moreover, some sound events share similar audio patterns, but convey a distinct meaning.

The presented model demonstrates that even a basic semantic description of the environment can help to improve sound event recognition. The implementation of the model is not conclusive. Some design choices can be revised, such as including inhibitory connections in the network to represent counter associations. However, the overall design choice for a flexible model instead of a model with conditional dependencies, such as hidden Markov models, is important. Models based on conditional dependencies assume an exhaustive knowledge of the problem domain. As a consequence, missing knowledge, for example because of an unreliable signal, has a major impact on the output decision. For example, if one event in a learned sequence of events is not observed, the complete sequence will have a low probability (Box 3.3). In contrast, the integrated approach presented in this thesis provides a best hypothesis at any point in time when information is segregated from the signal. Furthermore, this best hypothesis is the result of a balance between signal-driven information and an expectancy based on knowledge of the context. Therefore, salient signal information can override a falsely inferred expectancy.[1]

[1] However, parts of the Bayes formalism, such as learning a priori relations between events and environments, can be useful.

7.2 IMPLICATIONS

To further improve our method for environmental sound event recognition several issues have to be addressed, which are all related to selective attention. First, the signal-driven methods and the context model are not yet interactive. The context model interprets the results of the signal-driven method, but does not influence the search space of the signal analysis. In future versions the context model should influence the priority of the signal analysis. In other words, the segregation algorithms should attend to parts of the signal that are likely to provide relevant information given the context, and ignore irrelevant parts. For example, a continuous low frequency component caused by traffic noise is informative when it is unlikely in a certain context, like a natural park, while in an urban environment this component can often be ignored because it is normal for this context. However, context-based attention is not the only determinant for the signal analysis. If some patterns in the audio signal are salient, for example due to an increase in loudness, they should be processed regardless of the priority given by the context model.

A second improvement can be made in the evaluation method for sound event recognition systems. A system for counting cars at a road should recognize every passing car, while other information can be ignored. In contrast, if a system should evaluate a sonic environment in a similar way as (a group of) people, a different set of events is interesting or irrelevant. For example, most people walking by a busy road do not pay attention to every car driving by, while they may focus on other more interesting sounds.[1] The performance of a system for sound event recognition cannot be determined without a benchmark, which depends on the goal of the system. However, we do not want to limit our system to specific applications. As a consequence, the context model should be flexible enough to work in a variety of different applications. For every application the learning process and the benchmark are different. As a result, the context model attends to different parts of environment in different applications. The quality of the model can be determined by its performance in different domains, with different benchmarks.

Finally, the current implementation of the context model as a spreading acti-

[1] Which sound events are interesting depends on factors such as the goal or activity of a person, his memory, interpretation, and expectancy.

vation network can lead to computational issues in a diverse application domain, because of the large amount of different types of sound events that have to be managed in the network. However, even infinite capacity does not guarantee a system that can deal with a real-world environment. People effortlessly recognize and remember thousands of different types of sound events. Nevertheless, they do not process every piece of information they encounter, nor do they need to. By focussing their attention to parts of the environment, they make a selection of what is relevant. The fundamental difficulty of automatic systems in matching the effective strategies of people to process relevant information and act meaningfully upon it is called the frame problem (Dennett, 1990): "A walking encyclopedia will walk over a cliff, for all its knowledge of cliffs and the effect of gravity, unless it is designed in such a fashion that it can find the right bits of knowledge at the right times, so it can plan its engagements with the real world."

Even though the context model cannot learn all world knowledge and experience that people have, its focus can be narrowed by a goal, and it can learn which information is relevant given the goal and a particular context. In fact, instead of a focus on abundant learning, we want to apply knowledge about human cognition in the context model. People do not require tens or hundreds of examples to be able to recognize a sound event. They structure the world into categories, and new instances of an object or event are matched to prototypes (or exemplars) in memory. These prototypes are more than an average of the features of all members of a category. In cognitive psychology much research is aimed at understanding human categorization. We will try to translate the findings of this research into the context model. For example, if we understand how categorization is effected by a different context, or by the expectancy of a person, we can adjust the level of analysis of the model in a similar manner.

In summary, modeling information about the context is essential in automatic sound event recognition that should work in variable real-world environments. In this thesis we have substantiated the fundamental grounds for an integrated approach to sound event recognition, which combines robust signal-driven algorithms with a context model. Furthermore, we have demonstrated a first effort of an implementation of this integrated approach, which improves the performance results compared to an approach that is based on standard machine learning algorithms. A semantic analysis of a sonic environment that is obtained with a model

of human memory will have important applications in diverse fields. For example, instead of monitoring sound in an urban environment only with loudness measures, as is currently done, a semantic analysis of the sonic environment provides a richer account of human evaluation of the environment.

Appendix A: Glossary

audio component

Single element in the audio signal, which is (possibly together with other components) the result of a sound event. (p. 14)

audio pattern

Configuration of grouped audio components that constitute a single sound event. (p. 4)

audio signal

Signal transmitted by a physical action recorded or heard at the location of a receiver. (p. 2)

auditory episode

Cognitive representation of a sonic environment. (p. 28)

auditory event

Cognitive representation of a sound event. (p. 28)

context

The learned associations of an event to environments and/or co-occurring events. (p. 26)

context-based

When an algorithm or model processes knowledge of the context (cf. top-down). (p. 46)

physical action

Action or process of a source that results in a sound event. (p. 28)

segregate

Select and group components in the audio signal that are likely to constitute a single sound event. (p. 4)

semantic

When an interpretation or representation is based on the experienced properties of something and its relation to other things. It is similar to the meaning that can be attributed by people. However, it can be given by systems as well, whereas meaning can only be attributed by people. (p. 26)

signal-driven

When an algorithm or model processes the (audio) signal (cf. bottom-up). (p. 8)

sound

Sound that is not attributed to a specific event. (p. 2)

sound event

Sound that is the result of a single physical action. (p. 7)

sound source

Source involved in the physical action that produces a sound event. (p. 7)

APPENDIX B: AUDIO PROCESSING

The audio processing used in this thesis, which is developed and described by Krijnders (2010), is shortly summarized in this appendix, based on the description in Krijnders *et al.* (2010). The first step consists of the conversion of the audio signal in a time-frequency representation, the cochleogram. In the cochleogram, tones and pulses are selected based on their local properties, which can be connected in time or frequency to form signal components. Tonal signal components can be combined into harmonic complexes when they comply with certain properties. Finally, broadband events are defined and extracted as a time delimited energy increase compared to the background noise.

B.1 COCHLEOGRAM

The audio time signal is processed using a gammachirp filter bank (Irino and Patterson, 1997). The response of each gammatone filter g_t is calculated as

$$g_t(t) = a t^{N-1} \exp(-2\pi b B(f_c) t) \exp(j(2\pi f_c t + c \log_{10} t)) \tag{B.1}$$

where f_c is the center frequency of the channel, N the order of the gammatone ($N = 4$) and constants $a = 1$, $b = 0.71$, and $c = -3.7$. A logarithmic frequency distribution is used for 100 channels between 67 and 4000 Hertz (Hz). The bandwidth of each filter is given by (Moore and Glasberg, 1996):

$$B(f_c) = 24.7 + 0.108 f_c. \tag{B.2}$$

The filter output is squared and leaky-integrated with a segment-dependent time constant ($\tau_s = 2/f_c$). The resulting energy representation is down-sampled to 200 Hz, resulting in a frame size of 5 milliseconds. The energy is compressed logarithmically and expressed in decibel (dB). We call this representation a cochleogram (see for example Figure 2.1).

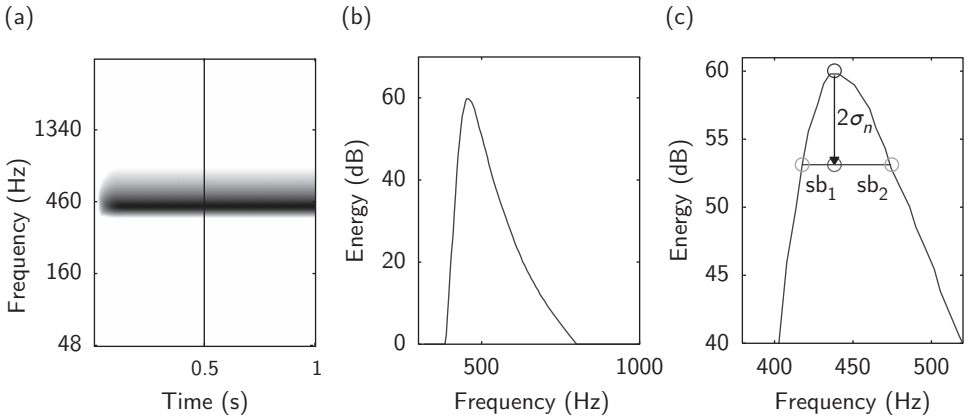


Figure B.1: Cochleogram of an ideal tone (a), the energy of the tone at 0.5 seconds (b) with filter parameters (c).

B.2 TONE FIT AND PULSE FIT

To extract tones and pulses from the cochleogram, we apply channel dependent matched filters that respond to ideal tones and pulses. For each channel an ideal tone is generated and processed using the filter bank. In Figure B.1, panel (a) shows the time-frequency representation of the tone, and panel (b) the energy of the tone in one time frame. The width of the response in frequency at a threshold below the energy maximum is calculated, depicted in panel (c). This threshold is set to twice the standard deviation of the logarithmic energy of white noise in the channel ($2\sigma_n$). This standard deviation is independent of the power spectral density of the noise in the logarithmic energy domain. The width of the response is the filter parameter for the tone fit (TF). For the pulse fit (PF) the response width in time of a pulse is taken.

When the filters are applied in a cochleogram, the energies at the widths below (sb_1) and above (sb_2) a time-frequency point are averaged. The difference between the energy at the point for that channel and the average forms the filter output (Figure B.2). The application of the filters to the cochleogram results in two representations that reflect to what extend the direct environment of each point of the cochleogram resembles a tone or a pulse. These representations are thresholded to create a binary mask. This threshold is set to twice the standard deviation of

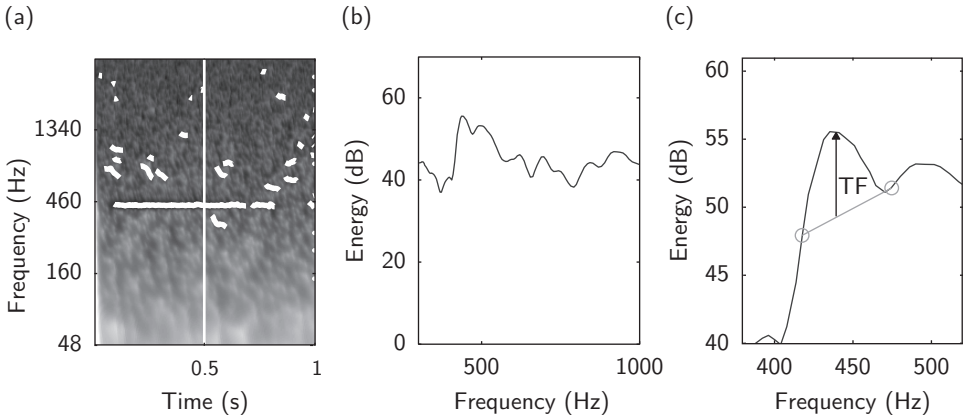


Figure B.2: TF filter applied to a tone in 0 dB local SNR white noise.

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 5.2).

If possible, the tonal signal components are combined into harmonic complexes (HCs). 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 Hz 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:

$$S = n_{sc} + b_{f0} + n_h - \sum_{sc} \mathrm{rms}_{sc} - \sum_{sc} \Delta f_{sc} \tag{B.3}$$

where n_{sc} is the number of signal components in the group, b_{f0} is a boolean for 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: The duration, score S (equation B.3), the ratio of these two, and the number of signal components indicate the strength of a harmonic complex. The mean energy and standard deviation under the signal components, the spectral tilt of the signal components, and the mean and standard deviation of the fundamental frequency are copied from Van Hengel and Andringa (2007) and Zajdel *et al.* (2007) to discriminate between similar harmonic sounds, such as speech and laughing.

B.3 BROADBAND EVENTS

Broadband events are defined as slow broadband changes in the signal that have to satisfy the following 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 a time interval. This level of 95% assumes that each channel is dominated by background noise at least 5% of the time interval with a temporal scope depending on the data set—the experiments described in chapter 5 use the length of the recordings, typically between one and three minutes. The energy must exceed the background by three standard deviations of white noise in that channel.

The broadband events 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. 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 third property is the average foreground-to-background ratio for each band, which reflects the total energy per band compared to the background.

APPENDIX C: REVERBERATION FEATURES

The seven features that indicate the reverberation level measure one of three types of fluctuation: variation in the energy of the harmonic tracks, variation in the salience of the harmonic tracks, and variation in the frequency of the harmonic tracks. The calculation of the features is described in the following sections. We apply the following notation: h is the harmonic track, extracted from a cochleogram (see appendix B), $E_h(t)$ is energy development of the harmonic track in decibels (dB), and $f_h(t)$ is the frequency development of the harmonic track in Hertz (Hz).

C.1 ENERGY VARIATION

Peak rate (PR) The number of peaks (energy maxima) of at least 1 dB in the energy of the harmonic track, E_h, normalized for the length of the track.

Var E_h Variation difference between the energy of the harmonic track and its moving average (normalized for the length of the track), calculated as

$$\text{Var } E_h = \sum_t \left| \mathrm{d}t\, E_h(t) \right| - \sum_t \left| \mathrm{d}t\, \overline{E}_h(t) \right|,$$

where $\mathrm{d}t\, E_h(t)$ is the differential of the harmonic track energy $E_h(t)$, and $\overline{E}_h(t)$ is its moving average:

$$\overline{E}_h(t) = \frac{1}{k} \sum_{i=t}^{t+k-1} E_h(i), \text{ for } t = 1 : n-k+1,$$

where E_h has a total length of n frames (1 frame is 5 milliseconds), and the applied window size is $k = 7$ frames.

C.2 HARMONIC ENERGY SALIENCE

$\Delta E_h(f)$ The energy slope is calculated as the mean difference in energy between a harmonic track and a reference track E_{ref} at a higher frequency and at a lower

frequency side of the harmonic track:

$$\Delta E_h(f) = \overline{E_h - E_{\text{ref}}},$$

where E_{ref} is determined by the difference between the fundamental frequency f_0 and a track at $k = 1.2$ or 0.8 of f_0 (the fundamental frequency is determined through the harmonic complex, see appendix B):

$$E_{\text{ref}}(f, t) = E(f_h(t) \pm |k f_0(t) - f_0(t)|, t), \text{ for } t = 1 : n,$$

where n is the length of the harmonic track and the fundamental frequency.

Var Δf_h Variation difference between the frequency width of the harmonic track and an ideal sinusoid (normalized for the length of the track), calculated as

$$\text{Var } \Delta f_h = \sum_t \left| \mathrm{d}t \, \Delta f_s(t) \right| - \sum_t \left| \mathrm{d}t \, \Delta f_h(t) \right|,$$

where $\mathrm{d}t \, \Delta f_s(t)$ is the differential of the width of an ideal sinusoid at the same frequency as the harmonic track at a particular time, and $\Delta f_h(t)$ is the width in frequency channels of the harmonic track at time t, determined by the tone fit (TF), a frequency-dependent filter based on the ideal tone response of the cochlea model:

$$\Delta f_h(t) = \frac{\theta_{\text{TF}} - \text{TF}(s_{\text{down}} + 1, t)}{\text{TF}(s_{\text{down}} + 1, t) - \text{TF}(s_{\text{down}}, t)} - \frac{\theta_{\text{TF}} - \text{TF}(s_{\text{up}} - 1, t)}{\text{TF}(s_{\text{up}}, t) - \text{TF}(s_{\text{up}} - 1, t)},$$

where $\theta_{\text{TF}} = 1.3$ is the threshold of the tone fit, $\text{TF}(s_{\text{down}}, t)$ is the response of the TF filter at time t in channel s_{down}, the channel at the high frequency end of the harmonic region[1], and s_{up} is the channel at the low frequency end of the harmonic region. This harmonic region is determined through a tone mask, a binary mask created by thresholding the TF filter response of the cochleogram at twice the standard deviation of the TF filter when applied to white noise (see appendix B).

[1] The channel numbers start at the high frequency side of the cochleogram and end at the low frequency side.

Mean Δf_h Mean difference between the frequency width of the harmonic track and an ideal sinusoid, calculated as

$$\text{Mean } \Delta f_h = \overline{(\Delta f_s - \Delta f_h)},$$

where $\Delta f_s(t)$ and $\Delta f_h(t)$ are calculated as above.

C.3 HARMONIC FREQUENCY SALIENCE

Var f_h/MA Variation difference between the frequency of the harmonic track and its moving average (normalized for the length of the track), calculated as

$$\text{Var } f_h = \sum_t \left| \mathrm{d}t\, f_h(t) \right| - \sum_t \left| \mathrm{d}t\, \overline{f}_h(t) \right|,$$

where $\mathrm{d}t\, f_h(t)$ is the differential of the harmonic track frequency $f_h(t)$, and $\overline{f}_h(t)$ is its the moving average:

$$\overline{f}_h(t) = \frac{1}{k} \sum_{i=t}^{t+k-1} f_h(i), \text{ for } t = 1 : n - k + 1,$$

where f_h has a total length of n frames (1 frame is 5 milliseconds), and the applied window size is $k = 7$ frames.

Var f_h/P Variation difference between the frequency of the harmonic track and its polynomial approximation (normalized for the length of the track), calculated as

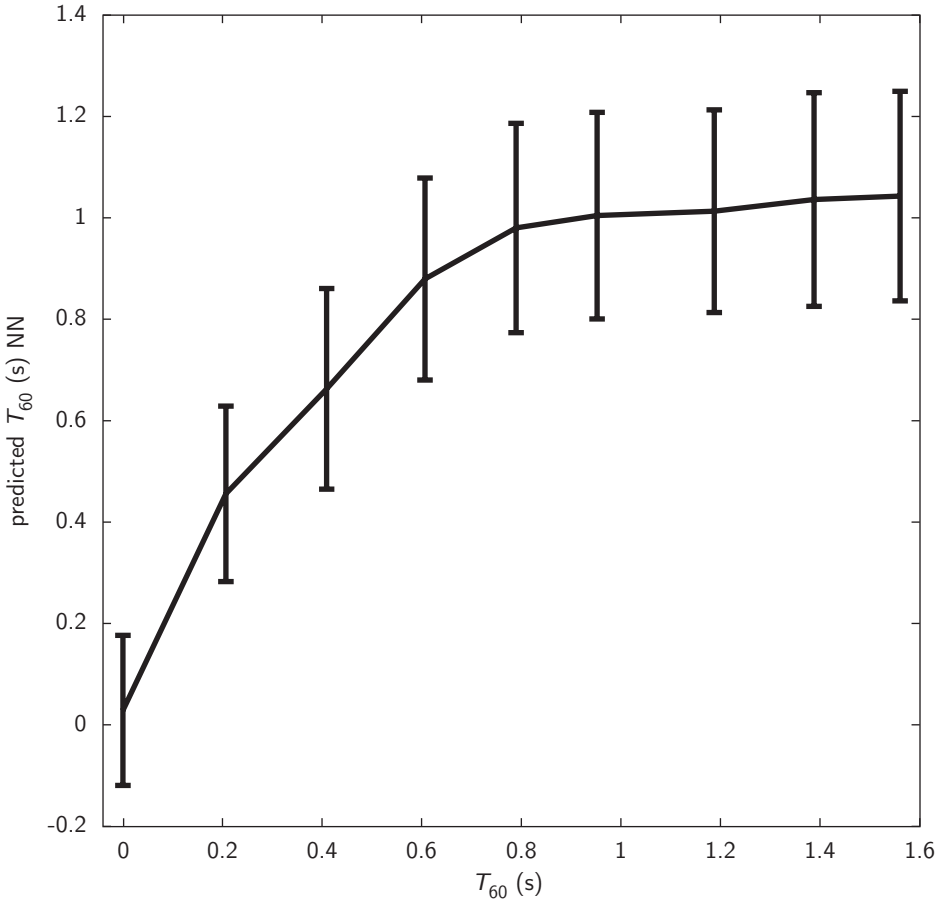
$$\text{Var } f_h = \sum_t \left| \mathrm{d}t\, f_h(t) \right| - \sum_t \left| \mathrm{d}t\, P(f_h(t)) \right|,$$

where $\mathrm{d}t\, f_h(t)$ is the differential of the harmonic track frequency $f_h(t)$, and $P(f_h(t))$ is its *piecewise cubic Hermite interpolating polynomial* (Fritsch and Carlson, 1980), calculated with five equally spaced time-frequency points on the track, that is, the start, the end, and three points in between.[1]

[1] This function is provided as *pchip* in Matlab (© 1984-2004 The MathWorks, Inc.).

C.4 PREDICTIVE STRENGTH OF FEATURES

A two-layer feed-forward backpropagation neural network (NN) was trained with the features on the data set described in section 2.4.2, of which 2595 samples were used for training, and 1297 samples for validating and testing, the results of which are depicted in the following figure:



REFERENCES

Adams, R. B. and Janata, P. (2002). A comparison of neural circuits underlying auditory and visual object categorization. *Neuroimage 16*(2), 361–377.

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America 65*(4), 943–950.

Anderson, J. R. (2005). *Cognitive Psychology and its Implications*. New York, NY: Worth Publishers, 6th ed.

Andringa, T. C. and Niessen, M. E. (2006). Real-world sound recognition: A recipe. In *Proceedings of the First International Workshop on Learning the Semantics of Audio Signals*, pp. 106–118.

Atienza, M., Cantero, J. L., Grau, C., Gomez, C., Dominguez-Marin, E., and Escera, C. (2003). Effects of temporal encoding on auditory object formation: A mismatch negativity study. *Cognitive Brain Research 16*(3), 359–371.

Aucouturier, J.-J., Defréville, B., and Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America 122*(2), 881–891.

Aucouturier, J.-J. and Pachet, F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research 32*(1), 83–93.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance 19*(2), 250–267.

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

Ballas, J. A. and Mullins, T. (1991). Effects of context on the identification of everyday sounds. *Human Performance 4*(3), 199–219.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience 5*(8), 617–629.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences 11*(7), 280–289.

Barker, J. P., Cooke, M. P., and Ellis, D. P. W. (2005). Decoding speech in the presence of other sources. *Speech Communication 45*(1), 5–25.

Bennett, M. R. and Hacker, P. M. S. (2001). Perception and memory in neuroscience: A conceptual analysis. *Progress in Neurobiology 65*(6), 499–543.

Blauert, J. (2001). *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, third ed.

Blauert, J. and Jekosch, U. (1997). Sound-quality evaluation: A multi-layered problem. *Acustica 83*(5), 747–753.

Brandstein, M. S. and Ward, D. B. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin Heidelberg New York: Springer-Verlag.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

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

Cai, R., Lu, L., and Hanjalic, A. (2008). Co-clustering for auditory scene categorization. *IEEE Transactions on Multimedia 10*(4), 596–606.

Cai, R., Lu, L., Hanjalic, A., Zhang, H. J., and Cai, L. H. (2006). A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing 14*(3), 1026–1039.

Carello, C., Anderson, K. L., and Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science 9*(3), 211–214.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences 8*(10), 465–471.

Carrell, T. D. and Opie, J. M. (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception & Psychophysics 52*(4), 437–445.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America 25*(5), 975–979.

Chu, S., Narayanan, S., and Kuo, C. . C. J. (2009). Environmental sound recognition with time-frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing 17*(6), 1142–58.

Cohen, P. R. and Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management 23*(4), 255–268.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review 82*(6), 407–428.

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

Côté, M., Lecolinet, E., Y., M. C. C., and Suen (1998). Automatic reading of cursive scripts using a reading model and perceptual concepts. *International Journal on Document Analysis and Recognition 1*(1), 3–17.

Couvreur, L. and Couvreur, C. (2004). Blind model selection for automatic speech recognition in reverberant environments. *Journal of VLSI Signal Processing 36*(2-3), 189–203.

Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters 24*(15), 2895–2907.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review 11*(6), 453–482.

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience 17*(4), 641–651.

Darwin, C. J. and Hukin, R. W. (2000). Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *The Journal of the Acoustical Society of America 108*(1), 335–342.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 28*(4), 357–366.

Defréville, B., Roy, P., Rosin, C., and Pachet, F. (2006). Automatic recognition of urban sound sources. In *Proceedings of the 120th AES Convention*.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1990). Cognitive wheels: The frame problem of AI. In M. A. Boden (Ed.), *The Philosophy of Artificial Intelligence*, pp. 147–170. Oxford University Press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy 88*(1), 27–51.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quartily 1*(1), 35–68.

Dubois, D., Guastavino, C., Maffiolo, V., and Raimbault, M. (2004). A cognitive approach

to soundscape research. *The Journal of the Acoustical Society of America 115*(5), 2495.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. New York, NY: John Wiley & Sons, second ed.

Dyson, B. J. and Alain, C. (2004). Representation of concurrent acoustic objects in primary auditory cortex. *The Journal of the Acoustical Society of America 115*(1), 280–288.

Ellis, D. P. W. (1996). *Prediction-Driven Computational Auditory Scene Analysis*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing 14*(1), 321–329.

Eyring, C. F. (1930). Reverberation time in "dead" rooms. *The Journal of the Acoustical Society of America 1*(2A), 168.

Fastl, H. (1997). The psychoacoustics of sound-quality evaluation. *Acustica 83*(5), 754–764.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition 28*(1-2), 3–71.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America 99*(3), 1730–1741.

Fritsch, F. N. and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis 17*(2), 238–246.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance 6*(1), 110–125.

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

Genuit, K. and Fiebig, A. (2006). Psychoacoustics and its benefit for the soundscape approach. *Acta Acustica United with Acustica 92*(6), 952–958.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin Company.

Godsmark, D. J. and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication 27*, 351–366.

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

Grossberg, S. (1980). How does a brain build a cognitive code. *Psychological Review 87*(1), 1–51.

Grossberg, S., Govindarajan, K. K., Wyse, L. L., and Cohen, M. A. (2004). ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks 17*(4), 511–536.

Guastavino, C. (2006). The ideal urban soundscape: Investigating the sound quality of French cities. *Acta Acustica United with Acustica 92*(6), 945–951.

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

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *The Journal of Neuroscience 25*(22), 5382–5388.

Gygi, B., Kidd, G. R., and Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & Psychophysics 69*(6), 839–855.

Gygi, B. and Shafiro, V. (2006). Effect of context on identification of environmental sounds. *The Journal of the Acoustical Society of America 119*(5), 3334.

Gygi, B. and Shafiro, V. (2007). General functions and specific applications of environmental sound research. *Frontiers in Bioscience 12*, 3152–3166.

Hansen, H. and Weber, R. (2009). Semantic evaluations of noise with tonal components in Japan, France, and Germany: A cross-cultural comparison. *The Journal of the Acoustical Society of America 125*(2), 850–862.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena 42*(1-3), 335–346.

Hirsch, H. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181–188.

Hollingworth, A. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General 127*(4), 398–415.

Irino, T. and Patterson, R. D. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America 101*(1), 412–419.

Jones, S. J., Longe, O., and Pato, M. V. (1998). Auditory evoked potentials to abrupt pitch and timbre change of complex tones: electrophysiological evidence of 'streaming'? *Evoked Potentials-Electroencephalography and Clinical Neurophysiology 108*(2), 131–142.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Models for speech recognition. *Technometrics 33*(3), 251–272.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing 24*(1), 1–10.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research 4*, 237–285.

Kinoshita, K., Delcroix, M., Nakatani, T., and Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing 17*(4), 1–12.

Köhler, W. (1967). Gestalt psychology. *Psychological Research 31*(1), XVIII–XXX.

Kootstra, G., Jong, S. D., and Schomaker, L. (2009). Using local symmetry for landmark selection. In *Computer Vision Systems: Proceedings of the 7th International Conference on Computer Vision Systems*, pp. 94–103.

Krijnders, J. D. (2010). *Signal-Driven Sound Processing for Uncontrolled Environments*. Ph.D. thesis, University of Groningen.

Krijnders, J. D., Niessen, M. E., and Andringa, T. C. (2007). Robust harmonic complex estimation in noise. In *Proceedings of the 19th International Congress on Acoustics*.

Krijnders, J. D., Niessen, M. E., and Andringa, T. C. (2010). Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters 31*(12), 1552–1559.

Kubovy, M. and Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition 80*(1-2), 97–126.

Kunkler-Peck, A. J. and Turvey, M. (2000). Hearing shape. *Journal of Experimental Psychology: Human Perception and Performance 26*(1), 279–294.

Kuttruff, H. (1979). *Room Acoustics*. London, UK: Applied Science Publishers, second ed.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Interna-*

tional Journal of Computer Vision 60(2), 91–110.

Lu, L. and Hanjalic, A. (2008). Audio keywords discovery for text-like audio content analysis and retrieval. *IEEE Transactions on Multimedia 10*(1), 74–85.

Lu, L. and Hanjalic, A. (2009). Text-like segmentation of general audio for content-based retrieval. *IEEE Transactions on Multimedia 11*(4), 658–669.

Marcell, M. E., Borella, D., Greene, M., Kerr, E., and Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology 22*(6), 830–864.

Maris, E., Stallen, P. J., Vermunt, R., and Steensma, H. (2007). Evaluating noise in social context: The effect of procedural unfairness on noise annoyance judgments. *The Journal of the Acoustical Society of America 122*(6), 3483–3494.

Marsland, S., Shapiro, J., and Nehmzow, U. (2002). A self-organising network that grows when required. *Neural Networks 15*(8-9), 1041–1058.

Matassoni, M., Omologo, M., and Giulini, D. (2000). Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1407–1410.

McAdams, S. (1993). Recognition of auditory sounds sources and events. In S. McAdams and E. Bigand (Eds.), *Thinking in Sound: The Cognitive Psychology of Human Audition*, pp. 146–198. Oxford University Press.

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

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review 85*(3), 207–238.

Moore, B. C. J. and Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica United with Acustica 82*(2), 335–345.

Nakatani, T., Juang, B. H., Yoshioka, T., Kinoshita, K., Delcroix, M., and Miyoshi, M. (2008). Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source models. *IEEE Transactions on Audio, Speech, and Language Processing 16*(8), 1512–1527.

Newell, A. (1982). The knowledge level. *Artificial Intelligence 18*(1), 87–127.

Niessen, M. E., Kootstra, G., de Jong, S., and Andringa, T. C. (2009). Expectancy-based

robot navigation through context evaluation. In *Proceedings of the 2009 Interenational Conference on Artificial Intelligence*, pp. 371–377.

Niessen, M. E., Krijnders, J. D., Boers, J., and Andringa, T. C. (2007). Assessing the reverberation level in speech. In *Proceedings of the 19th International Congress on Acoustics*.

Niessen, M. E., van Maanen, L., and Andringa, T. C. (2008*a*). Disambiguating sound through context. *International Journal on Semantic Computing 2*(3), 327–341.

Niessen, M. E., van Maanen, L., and Andringa, T. C. (2008*b*). Disambiguating sounds through context. In *Proceedings of the Second IEEE International Conference on Semantic Computing*, pp. 88–95.

Nix, J. and Hohmann, V. (2007). Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Transactions on Audio, Speech, and Language Processing 15*(3), 995–1008.

Norris, R. F. and Andree, C. A. (1929). An instrumental method of reverberation measurement. *The Journal of the Acoustical Society of America 1*(1), 32.

Nosofsky, R. A. and Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition 28*(5), 924–940.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences 11*(12), 520–527.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition 41*(10), 2965–2979.

Palmer, S. E. (1975). Effects of contextual scenes on identification of objects. *Memory & Cognition 3*(5), 519–526.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic Information Processing*, pp. 216–270. Cambridge, MA: MIT Press.

Radlovic, B. D., Williamson, R. C., and Kennedy, R. A. (2000). Equalization in an acoustic reverberant environment: Robustness results. *IEEE Transactions on Speech and Audio Processing 8*(3), 311–319.

Raimbault, M., Lavandier, C., and Berengier, M. (2003). Ambient sound assessment of urban environments: field studies in two French cities. *Applied Acoustics 64*(12), 1241–1256.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology 3*(3), 382–

407.

Roman, N. and Wang, D. (2006). Pitch-based monaural segregation of reverberant speech. *The Journal of the Acoustical Society of America 120*(1), 458–469.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General 104*(3), 192–233.

Rosenblum, L. D. (2004). Perceiving articulatory events: Lessons for an ecological psychoacoustics. In J. G. Neuhoff (Ed.), *Ecological Psychoacoustics*, pp. 219–248. Elsevier Academic Press.

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

Samuel, A. (1996). Phoneme restoration. *Language and Cognitive Processes 11*(6), 647–653.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication 49*(5), 336–347.

Schermerhorn, P. and Scheutz, M. (2009). The utility of affect in the selection of actions and goals under real-world constraints. In *Proceedings of the 2009 International Conference on Artificial Intelligence*.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition 80*(1-2), 1–46.

Schulte-Fortkamp, B. and Fiebig, A. (2006). Soundscape analysis in a residential area: An evaluation of noise and people's mind. *Acta Acustica United with Acustica 92*(6), 875–880.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417–425.

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

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science 270*(5234), 303–304.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences 12*(5), 182–186.

Smith, J. D. and Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*(1), 3–27.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature 416*, 87–90.

Sternberg, S. (1966). High-speed scanning in human memory. *Science 153*(3736), 652–654.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*(236), 433–460.

Tzanetakis, G. and Cook, P. (1999). Multifeature audio segmentation for browsing and annotation. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pp. 103–106.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing 10*(5), 293–302.

Van Hengel, P. W. J. and Andringa, T. C. (2007). Verbal aggression detection in complex social environments. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 15–20.

Van Maanen, L. (2007). Mediating expert knowledge and visitor interest in art work recommendation. In *Proceedings of the Workshop Lernen-Wissen-Adaption*, pp. 367–372.

Van Maanen, L., Van Rijn, H., van Grootel, M., Kemna, S., Klomp, M., and Scholtens, E. (2010). Personal Publication Assistant: Abstract recommendations by a cognitive models. *Cognitive Systems Research 11*(1), 120–129.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Ph.D. thesis, Eindhoven University of Technology.

Van Rijsbergen, C. J. (1979). In *Information Retrieval*, pp. 112–140. London, UK: Butterworths.

Van Valkenburg, D. and Kubovy, M. (2004). From gibson's fire to gestalts: A bridge-building theory of perceptual objecthood. In J. G. Neuhoff (Ed.), *Ecological Psychoacoustics*, pp. 113–147. Elsevier Academic Press.

Vanderveer, N. J. (1979). *Ecological acoustics: Human perception of environmental sounds*. Ph.D. thesis, Cornell University.

Vasudevan, S., Gächter, S., Nguyen, V., and Siegwart, R. (2007). Cognitive maps for mobile robots—an object based approach. *Robotics and Autonomous Systems 55*(5), 359–371.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis*. Holoken, NJ: John Wiley and Sons.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science 167*(3917), 392–393.

Warren, W. H. and Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance 10*(5), 704–712.

Winkler, I., van Zuijen, T. L., Sussman, E., Horváth, J., and Näätänen, R. (2006). Object representation in the human auditory system. *European Journal of Neuroscience 24*(2), 625–634.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann.

Wu, M. and Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Speech and Audio Processing 14*(3), 774–784.

Yost, W. A. (1991). Auditory image perception and analysis: The basis for hearing. *Hearing Research 56*(1-2), 8–18.

Yu, L. and Kang, J. (2008). Effects of social, demographical and behavioral factors on the sound level evaluation in urban open spaces. *The Journal of the Acoustical Society of America 123*(2), 772–783.

Zajdel, W., Krijnders, J. D., Andringa, T. C., and Gavrila, D. (2007). CASSANDRA: Audio-video sensor fusion for aggression detection. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 200–205.

Zhang, M. and Kang, J. (2007). Towards the evaluation, description, and creation of soundscapes in urban open spaces. *Environment and Planning B-Planning & Design 34*(1), 68–86.

PUBLICATIONS

Andringa, T. C. and Niessen, M. E. (2006). Real-world sound recognition: A recipe. In *Proceedings of the 1st International Workshop on Learning the Semantics of Audio Signals*, pp. 106–118.

Krijnders, J. D., Niessen, M. E., and Andringa, T. C. (2010). Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters 31*(12), 1552–1559.

Maisonneuve, N., Stevens, M., Niessen, M. E., Hanappe, P., and Steels, L. (2009*a*). Citizen noise pollution monitoring. In *Proceedings of the 10th Annual International Conference on Digital Government Research, Partnerships for Public Innovation*, pp. 96–103.

Maisonneuve, N., Stevens, M., Niessen, M. E., and Steels, L. (2009*b*). Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*, pp. 215–228.

Niessen, M. E., Kootstra, G., de Jong, S., and Andringa, T. C. (2009*a*). Expectancy-based robot navigation through context evaluation. In *Proceedings of the 2009 Interenational Conference on Artificial Intelligence*, pp. 371–377.

Niessen, M. E., Krijnders, J. D., and Andringa, T. C. (2009*b*). Understanding a soundscape through its components. In *Proceedings of Euronoise*.

Niessen, M. E., Krijnders, J. D., Boers, J., and Andringa, T. C. (2007). Assessing the reverberation level in speech. In *Proceedings of the 19th International Congress on Acoustics*.

Niessen, M. E., van Maanen, L., and Andringa, T. C. (2008*a*). Disambiguating sound through context. *International Journal on Semantic Computing 2*(3), 327–341.

Niessen, M. E., van Maanen, L., and Andringa, T. C. (2008*b*). Disambiguating sounds through context. In *Proceedings of the 2nd IEEE International Conference on Semantic Computing*, pp. 88–95.

Sprenger, S., Van Rijn, H., Niessen, M. E., and Van Maanen, L. (in preparation). Dutch norms for name agreement in environmental sound naming.

SUMMARY

The research domain of automatic sound event recognition aims to describe an audio signal in terms of the sound events that compose a sonic environment. The ability to recognize events in a real-world environment requires a listener or system to separate the sound events from each other and the background. Furthermore, these separated events need to be recognized. To recognize a sound event implies that some representation of the event is already known to the receiver, and can be identified when it is encountered again. The ability of sound event recognition depends not only on the audio pattern specific to the event, but also on the semantics of the event. For example, a sound of a purring cat may seem unique, but without any other information than represented by the audio signal, it can sound like an engine as well. In this thesis we show that the task of recognizing a sound event can be alleviated with the semantics of the event, which is inferred from a model of the context in which the event occurs.

A possible strategy to provide an automatic sound recognition system with the semantics of a sound event, is to develop it for a specific application, and hence a specific type of context. For example, automatic speech recognition systems expect a speech signal as input, which is ensured by a user. Therefore, particular assumptions about the audio signal can be made, and context information in the form of grammar rules can be applied to recognize a word sequence (section 2.2). However, if a system for automatic sound event recognition is not designed for a specific application, and should work in variable and uncontrolled real-world environments, no assumptions about the environmental conditions can be made. Therefore, additional analysis of the audio signal is required. First, the sound events to be recognized have to be separated from the background, because the input signal consists of more than one type of sound (section 2.3). Second, the operating environment cannot be controlled, hence the system has to deal with transmission effects, such as reverberation (section 2.4).

In addition to handling the challenges of real-world environments with signal-driven methods, the semantics of the event and its environment are essential to recognize sound events in an unreliable or ambiguous audio signal. People have no difficulty in recognizing sound events in many different and noisy situations. Hence, we developed a model for automatic sound event recognition that is inspired by the strategies of human listeners as investigated by (psycho-) acoustics and cognitive psychology. The human percept of a sound event, referred to as an auditory event (section 3.2), has properties that a representation of a sound event in an automatic system can benefit from as well. People can generalize auditory events over different experiences, environments, and senses. Therefore, our model should store invariant representations of sound events, so that it is robust to variable environmental conditions, similar to human perception. Moreover, people benefit from information about the environmental context to recognize sound events (section 3.3). This facilitatory effect of contextual knowledge is an important design objective of our model.

Some other studies have been aimed at modeling context awareness in acoustics (section 4.1). However, in these studies the goal has been to estimate the context in itself, rather than to use context for the improvement of sound event recognition. In other research domains, such as information retrieval and handwriting recognition, context has been used to improve recognition or retrieval of objects. Often methods in these research domains use spreading activation networks, which are based on a model of human memory (Collins and Loftus, 1975). In this thesis we show that spreading activation networks can also be applied to estimate the most likely interpretation of an audio pattern. We introduce a context model in which the semantics of the events and the context are represented as nodes in the network (section 4.2). The activation of the nodes spreads through the network to determine the confidence of possible interpretations of an audio pattern (section 4.3).

The advantage of modeling context in automatic sound event recognition is demonstrated by applying an integrated system to audio recorded in real-world environments. This integrated system is a combination of a signal-driven analysis of the audio signal, which provides hypotheses of sound events, and the context model, which interprets these hypotheses. Knowledge about the environmental context is learned in a training phase. This knowledge is represented as a network

of nodes, in which the nodes represent the semantics of the sound events and the different contexts. Furthermore, the connections between the nodes carry weights that indicate the probabilities that these sound events and contexts are encountered subsequently or concurrently. The values of the weights between nodes are learned from annotated training data (section 5.2). The type of context that is learned depends on the application domain and the data set. In a stable environment, information of co-occurring events can help to form expectancies of future events. For example, at a train station the beeping of a closing door is likely to be followed by a departing train. Alternatively, if the data set is recorded at qualitatively different types of locations, the estimated location can help to predict the types of sound events that may be heard. For example, birds are more commonly heard in parks than near busy roads. We explored the benefit of both types of context on sound event recognition in two experiments (section 5.3 and 5.4). The results of these experiments show that the evaluation of contextual knowledge improves the recognition of sound events compared to an exclusively signal-driven method.

Contextual knowledge is not restricted to knowledge that can be derived from the audio signal and annotations of the audio signal. Other types of knowledge can be beneficial for sound event recognition as well. Knowledge inferred from different types of input, such as sound, image, and location, can reinforce each other to obtain an increased awareness of events in the environment. Ideally, these different informational resources can be combined in a single system. Because the nodes in the context model are not described by modality specific knowledge, they can be used for other types of information. We tested the applicability of the context model on the recognition of ambiguous visual information, in the domain of robot localization. Visual information received by a robot is often ambiguous (similar to acoustic information), because similar observations, such as (parts of) chairs or windows, can be made at distinct places in an environment. Learned knowledge about the environment, and the robot's hypothesized position in the environment, can help to disambiguate these observations. As a result, the position prediction improves compared to a signal-driven approach (chapter 6).

SAMENVATTING

Het onderzoeksgebied automatische geluidsherkenning richt zich op het beschrijven van een geluidssignaal in termen van de bronnen die het opgenomen geluid veroorzaken. Het vermogen om de bron van een geluid te herkennen vereist dat een luisteraar of systeem verschillende geluiden van elkaar en van de achtergrond kan scheiden. Bovendien moeten deze gescheiden geluiden herkend worden. Het herkennen van een geluid veronderstelt dat een representatie van het geluid al bekend is bij de luisteraar en geïdentificeerd kan worden wanneer deze opnieuw wordt waargenomen. Het herkennen van een geluid hangt niet alleen af van de kenmerken van het specifieke geluidspatroon, maar ook van de betekenis van de gebeurtenis die het geluid veroorzaakt. Het geluid van een spinnende kat bijvoorbeeld, mag wellicht uniek lijken, maar zonder enige informatie anders dan wat afgeleid kan worden uit het geluidssignaal, kan het evengoed als een motor klinken. In dit proefschrift laten we zien dat geluidsherkenning vereenvoudigd kan worden door gebruik te maken van de betekenis van de gebeurtenis die het geluid veroorzaakt. Deze betekenis wordt afgeleid middels een model dat de context van de geluidsomgeving representeert.

Een mogelijke manier om een automatische geluidsherkenner te voorzien van de betekenis van een geluid, is door het te ontwikkelen voor een specifieke applicatie, en derhalve voor een specifieke context. Automatische spraakherkenning verwacht bijvoorbeeld een spraaksignaal als invoer, wat door een gebruiker wordt gegarandeerd. Hierdoor kunnen bepaalde aannames over het geluidssignaal gemaakt worden, en contextinformatie in de vorm van grammaticaregels kan toegepast worden om een woordenreeks te herkennen (sectie 2.2). Als een system voor automatische geluidsherkenning echter niet ontworpen is voor een specifieke applicatie en moet werken in wisselende en ongecontroleerde omgevingen, kunnen er geen aannames worden gemaakt over de omgevingscondities. Daarom is aanvullende analyse van het geluidssignaal nodig. Ten eerste moeten de individuele

geluiden van elkaar en van de achtergrond worden gescheiden, omdat het invoersignaal bestaat uit meer dan een type geluid (sectie 2.3). Ten tweede kan de toepassingsomgeving niet gecontroleerd worden, waardoor het systeem moet omgaan met transmissieverschijnselen zoals galm (sectie 2.4).

In aanvulling op de signaalgedreven methoden om de uitdaging van ongecontroleerde omgevingen te beheersen, zijn de betekenis van de gebeurtenis en de geluidsomgeving essentieel om geluiden te herkennen in onbetrouwbare of ambigue geluidssignalen. Mensen hebben geen probleem om geluiden te herkennen in veel verschillende types omgevingen en in zeer rumoerige omgevingen. Daarom hebben we een model ontwikkeld voor automatische geluidsherkenning dat ingegeven is door de strategieën die mensen gebruiken, onderzocht door psychoakoestiek en cognitieve psychologie. Menselijke waarneming van een geluid, aangeduid als een auditieve gebeurtenis (sectie 3.2), heeft eigenschappen die ook gunstig kunnen zijn voor een geluidsrepresentatie in een automatisch systeem. Mensen kunnen auditieve gebeurtenissen generaliseren over verschillende omgevingen, ervaringen, en zintuigen. Op een vergelijkbare manier kan ons model invariante representaties van geluiden opslaan, zodat het robuust is voor wisselende omgevingsinvloeden, net als in menselijke perceptie. Mensen halen bovendien voordeel uit de omgevingscontext om geluiden te herkennen (sectie 3.3). Dit faciliterende effect van contextinformatie is een belangrijke doelstelling voor de ontwikkeling van ons model.

Verscheidene andere studies in de akoestiek zijn gericht op het modeleren van contextbesef (sectie 4.1). In deze studies was het doel echter om de context zelf te bepalen, in plaats van de context te gebruiken om geluidsherkenning te verbeteren. In andere onderzoeksgebieden, zoals *information retrieval* en handschriftherkenning, is context wel gebruikt om herkenning of retrieval van objecten te bevorderen. Methoden in deze onderzoeksgebieden maken vaak gebruik van *spreading activation* netwerken, die gebaseerd zijn op een model van het menselijk geheugen (Collins and Loftus, 1975). In dit proefschrift laten we zien dat spreading activation netwerken ook toegepast kunnen worden om de meest waarschijnlijke interpretatie van een geluidspatroon te bepalen. We introduceren een contextmodel waarin de betekenissen van de gebeurtenissen en de context gerepresenteerd zijn als knopen in het netwerk (sectie 4.2). De activatie van de knopen verspreidt zich door het netwerk om de zekerheid van mogelijke interpretaties van een geluidspatroon

te bepalen (sectie 4.3).

Het voordeel van het modeleren van context in automatische geluidsherkenning wordt gedemonstreerd door een geïntegreerd systeem toe te passen op geluid dat is opgenomen in ongecontroleerde omgevingen. Dit geïntegreerde systeem is een combinatie van een signaalgedreven analyse van het geluidssignaal, die hypotheses levert over de gebeurtenissen die de geluiden veroorzaken, en het contextmodel, dat deze hypotheses interpreteert. Kennis over de omgevingscontext wordt geleerd in een trainfase. Deze kennis wordt gerepresenteerd als een netwerk van knopen, waarin de knopen de betekenis van de geluiden en de verschillende contexten representeren. De verbindingen tussen de knopen dragen gewichten die de waarschijnlijkheden van het simultaan of opeenvolgend voorkomen van de geluiden en contexten representeren. De waarden van de gewichten tussen de knopen worden geleerd uit geannoteerde traindata (sectie 5.2). Het type context dat wordt geleerd is afhankelijk van het toepassingsgebied en van de dataset. In een stabiele omgeving kan informatie over gebeurtenissen die gezamelijk voorkomen helpen om verwachtingen van toekomstige gebeurtenissen te vormen. Op een station wordt het fluiten van sluitende deuren bijvoorbeeld gewoonlijk gevolgd door een vertrekkende trein. Een andere mogelijkheid is dat wanneer de traindata zijn opgenomen in kwalitatief verschillende omgevingen, de ingeschatte omgeving kan helpen om te voorspellen welk type geluiden er gehoord kunnen worden. Vogels worden vaker gehoord in parken dan bij drukke straten bijvoorbeeld. We hebben het voordeel van beide typen context voor geluidsherkenning verkend in twee experimenten (sectie 5.3 en 5.4). De resultaten van deze experimenten laten zien dat de evaluatie van contextinformatie geluidsherkenning verbetert ten opzichte van een signaalgedreven methode.

Contextinformatie is niet beperkt tot informatie die afgeleid kan worden uit het geluidssignaal en annotaties van het geluidssignaal. Andere informatiebronnen kunnen ook gunstig zijn voor geluidsherkenning. Informatie afgeleid uit verschillende types invoer, zoals geluid, beeld en locatie, kunnen elkaar versterken om een groter besef van de gebeurtenissen in de omgeving te verkrijgen. Idealiter kunnen deze verschillende informatiebronnen gecombineerd worden in één systeem. Omdat de knopen in het contextmodel niet beschreven worden door informatie die specifiek is voor een modaliteit, kunnen ze ook gebruikt worden voor andere soorten informatie. We hebben de toepasbaarheid van het contextmodel

op ambigue visuele informatie getest in het onderzoeksdomein van robotlocalisatie. De visuele informatie die een robot ontvangt is vaak ambigu (vergelijkbaar met akoestische informatie), omdat overeenkomstige observaties, zoals (delen van) stoelen of ramen, gemaakt kunnen worden op meerdere plekken in een omgeving. Geleerde informatie over de omgeving en de geschatte positie van de robot in die omgeving kan helpen om deze observaties ondubbelzinnig te maken. Hierdoor verbetert de voorspelling over de positie vergeleken met een signaalgedreven methode (hoofdstuk 6).

Dankwoord

Allereerst wil ik Tjeerd Andringa bedanken, die mijn project heeft opgestart en begeleid. Je hebt me de vrijheid gegeven om de richting van mijn onderzoek te bepalen, maar was wel altijd beschikbaar voor een gesprek. In die gesprekken verkenden we de mogelijkheden van mijn onderzoek, niet de beperkingen, en ik haalde daar veel inspiratie uit. Verder heb ik ook veel gehad aan de hulp van mijn promotor, Lambert Schomaker. Ondanks het verschil in vakgebied zijn je wetenschappelijke ervaring en je kennis van de kunstmatige intelligentie een waardevolle aanvulling geweest.

Naast de begeleiders die direct betrokken waren bij mijn onderzoek, heb ik natuurlijk ook veel steun gekregen van andere collega's. Ronald, als mijn kamergenoot was je niet alleen beschikbaar voor nuttige feedback op teksten of ideeën, maar ook voor een gezellig gesprek over van alles en nog wat. En niet te onderschatten, dankzij jouw verse koffie was ik niet afhankelijk van de machines! Ook de andere collega's uit onze vakgroep, Bea, Dirkjan, Hedde, en Renante, op de andere kamer, jullie zijn zowel in de samenwerking als in gezelligheid belangrijk geweest. We hebben veel kritische en leuke discussies gehad, gelukkig niet alleen op (en over) het werk, maar ook tijdens een barbecue of in de Minnaar.

Een aantal collega's met wie ik samen aan een project heb gewerkt wil ik bedanken voor de goede en prettige samenwerking. Dirkjan's project is nauw verbonden geweest met mijn project, en ons gecombineerd onderzoek heeft een meerwaarde gecreëerd. Onze samenwerking is altijd heel soepel verlopen, en ik ben blij dat we die kunnen voortzetten. Leendert, door ons project ben ik geïnteresseerd geraakt in cognitieve psychologie. Jouw ervaring en kennis in dit vakgebied toegepast in ons project zijn bepalend geweest voor de richting van mijn promotie. En Gert, dank je voor het voorstel om onze onderzoeksonderwerpen te combineren. We hebben een paar avonden moeten doortrekken, maar het (uiteindelijk) goede resultaat en het biertje na afloop maakten het zeker waard.

Alle overige collega's bij KI, Anja, Arnold, Axel, Bart, Ben, Elske, Esther, Fokie, Geertje, Hanneke, Hedderik, Ingrid, Jacolien, Jelmer, Jennifer, Jolie, Marco, Margriet, Mariëtte, Marius, Nancy, Niels, Petra, Rineke, Ronald, Roy, Sietse, Sietse, Sjoerd, Sonja, Sujata, en Tijn, bedankt voor de samenwerking en voor de gezelligheid tijdens lunches en borrels. Specifiek wil ik nog het secretariaat, en met name Elina bedanken, voor de praktische hulp door de jaren heen, en vooral met alle regeldingen voordat ik daadwerkelijk dat diploma heb. Jullie snelle en goede werk maakt een verschil.

Voordat het proefschrift naar de beoordelingscommissie is gegaan hebben een aantal collega's de moeite genomen om delen van mijn proefschrift te lezen. Caroline, Ronald, Hedde, en Bea, bedankt voor jullie nuttige suggesties. Ook wil ik de beoordelingscommissie, bestaande uit Danièle Dubois, Jaap van den Herik, en Barbara Shinn-Cunningham, bedanken voor de tijd die ze hebben genomen voor het lezen en voor het nuttige commentaar.

Mijn nieuwe werkgever, INCAS^3, wil ik bedanken voor de mogelijkheden en het vertrouwen die ze me al gegeven hebben voordat ik een afgerond onderzoek kon laten zien. Bovendien heb ik dankzij hen van Deborah's vakmanschap en inspiratie gebruik kunnen maken voor de omslag foto. Deborah, ik ben erg blij met het resultaat!

Gert en Manon, fijn dat ik jullie naast me heb staan als paranimfen en bedankt voor de hulp met de organisatie van de dag! Gert, je was eerst een gezellige collega, maar nu een goede vriend, en Manon, wat ben je niet geweest, buurvrouw, studiegenoot, collega. Het is maar goed dat we ook vrienden zijn.

Bij frisbee heb ik veel plezier en ontspanning gevonden. Iedereen bij GD, bedankt voor de leuke frisbeejaren!

En als laatste, vrienden, familie en Niklas, bedankt voor de steun tijdens de promotie, maar vooral voor alles daarbuiten, omdat het leven nou eenmaal meer is dan werk.