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 (Schmerhorn 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.
CHAPTER 3

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”.

28
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 transferred 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)

<table>
<thead>
<tr>
<th>Properties of an auditory event:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Based on information from sound events</td>
</tr>
<tr>
<td>2. Separated from information about the rest of the sonic environment</td>
</tr>
<tr>
<td>3. Generalized over different experiences</td>
</tr>
<tr>
<td>4. Generalized over different senses</td>
</tr>
</tbody>
</table>

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 focuses 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 focuses 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 rhythms (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.

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.
CHAPTER 3

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). 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”.

36
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).

<table>
<thead>
<tr>
<th>Sound event</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pouring water</td>
<td>Refrigerator door</td>
</tr>
<tr>
<td>Thunder</td>
<td>Rain</td>
</tr>
<tr>
<td>Whistle</td>
<td>Football cheering</td>
</tr>
<tr>
<td>Footstep</td>
<td>Drum</td>
</tr>
<tr>
<td>Toilet flush</td>
<td>Rain</td>
</tr>
<tr>
<td>Meowing cat</td>
<td>Crying baby</td>
</tr>
<tr>
<td>Coughing</td>
<td>Barking dog</td>
</tr>
<tr>
<td>Basketball</td>
<td>Closing door</td>
</tr>
<tr>
<td>Ticking clock</td>
<td>Pingpong ball</td>
</tr>
<tr>
<td>Water bubbles</td>
<td>Horse running</td>
</tr>
<tr>
<td>Bowling</td>
<td>Thunder</td>
</tr>
<tr>
<td>Zipper</td>
<td>Car starting</td>
</tr>
</tbody>
</table>

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 “heeeeeeelp”, 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.