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. 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 focusing 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.