Conclusion

In this work I have argued that new signal descriptors for environmental sounds, i.e. all sounds that occur in our everyday life, are needed because currently used descriptors, for example Mel Frequency Cepstral Coefficients, pose assumptions on the signal that are unjustifiable for these sounds. For example the assumptions that other sources can be considered as small perturbations on the target and that the fine structure of the spectrum is unimportant, are unjustifiable for these sounds. Another difference in the approach is to base the identification of sound events on auditory objects, here defined as combinations of sonic evidence that can justifiable be selected from the signal as coming from a single event.

I have developed two measures, the tone-fit and pulse-fit, which indicate how similar the time-frequency plane is to the local energy distribution of a tone, respectively a pulse. These measures capture the smallest structure in a spectrum and the probability of perturbations due to other sources is small because of the spectro-temporal limited scope of the measures.

Both measures have properties that are beneficial for sound recognition systems in unconstrained environment. These are:

1. A well defined relation to the local signal-to-noise ratio in the range between -5 dB and 25 dB, which corresponds to a hardly audible tone in noise to a very distinct tone in minimal noise.

2. Accurate in measuring frequency (< 0.8%) and time (≈ 2ms).
3. Capable of separating tones close in frequency and pulses close in time.

4. Signal level independent.

5. Correlated with perceptual descriptions of real-world recordings.

These properties are valid if the following assumptions are satisfied. These assumptions are that the time-frequency plane locally (around a time-frequency point) is similar to the energy distribution of a tone, respectively a pulse. While these similarities can occur in noise, it is unlikely that these persist longer than a specific duration or span more than a specific frequency range. The associated threshold durations and ranges can be calculated from the properties of broadband signals. This allows to check that the assumptions are satisfied.

To form auditory objects, areas of the time-frequency plane that adhere to the assumptions are extracted. The tone-like and pulse-like regions are then abstracted and interpreted as tones, respectively pulses. I have termed these narrow signal components because they represent signal components that are narrow in frequency, respectively time. If appropriate the tonal components are grouped to form harmonic complexes. Broadband components are formed by the difference between two background models with different time-constants. The background models are not updated with in time-frequency areas where narrow signal components are present. These different, both narrow and broadband, components cover most sound events.

The formed objects are identified using standard, simple classification techniques. Different kinds of objects result in different feature vectors. Consequently, the incoming signal determines which type of features are extracted. These feature-vectors are classified using a nearest-neighbour classifier. To measure the performance of the classifier I have used the f-measure from information retrieval where every second is treated as a retrieval, this has the advantage that the number of non-targets is not required for the measure.

To evaluate the presented methods in terms of recognition performance I have gathered two datasets, one with 40 recordings of about 2 minutes with acted scenes on a busy Amsterdam train station and one with multiple recordings on five, qualitatively different, places at different times in a city environment. The first dataset (the “Amstel dataset”) is fully annotated by the author, in the second dataset (the “Assen dataset”) each set of locations is annotated by the students who made them. An analysis of the annotation process has made clear that human annotation is far from uniform and thus its results are not easily used as a golden standard to measure our performance, but it is the best we have to score recognition system for unconstrained, environmental sounds.

Performance on the datasets ranges from an f-measure of 0.02 for rare classes to 0.5 for common classes. Though not directly comparable, the inter-
annotator f-measure on the “Assen dataset” was 0.46. This indicates that human annotators disagree on the annotations that form the ground truth with the same order as the system (dis)agrees with that ground truth.

To include context in the recognition system the recognition system was combined with a dynamic network model, developed in our group, that modeled co-occurrences and a larger context. This addition increases the f-measure on average 20% for the “Amstel dataset” and 33% for the “Assen dataset”.

To show the applicability of the recognition system, it was tested in several research projects. First, the harmonic complex extraction was extended to estimate formants in noisy speech and use these formants to recognize vowels. First, the recognition was combined with a video recognition system on the “Amstel dataset” for which also video recordings are available. Both recognizers fed a Bayesian network with the goal to detect aggression. We showed that the combining both modalities increased the detection rate (to 80%) while not increasing the number of false positives (16 per hour).

9.1 Future work

9.1.1 Signal Processing

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

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

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

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

9.1.2 Annotation

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

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

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

Apart from a ground truth, the annotation process can provide insight in how people listen to audio recordings. The full-temporal annotation is a very precise and analytical task which may provide the most complete annotation. However this is not a natural listening mode for humans. The “mid-point” annotation allows for less analytical, more natural listening and that may result in missing sound sources. If time-pressure is added more sources will go unnoticed. This may provide insights in what people judge to be important sources. This last qualification may also depend elements in the task description. For example, does changing the location mentioned in the task, change which sources are deemed important. The same factors should play a role in a automatic sound source recognition system, the system should only invest resources in analyzing the sound to the “start/stop” detail if the task of the system warrants it.
9.1.3 Recognition

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

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

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

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