Summary

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

Existing techniques for sound recognition are designed to function in closed, specialized domains. Speech recognition and music genre recognition, for example, work under the condition that the input is what they expect; speech from the speaker and the environment the system was trained on, or clean music recordings respectively. The idea that these closed domain techniques will generalize to open environments has so far not materialized. To operate in open environments we need to focus on the constancy and invariants in the signal: the physics that produced it.
In contrast to current “engineered”, specific systems, we aim to develop signal processing techniques that can handle sound in uncontrolled environments. Such environments are outside the range of the problems solved by current techniques, but are the normal environment for humans. These novel techniques are based on the research questions: “How to select sonic evidence that is likely to stem from a single source from a sound signal recorded in realistic acoustical circumstances?” and “How can the signal, instead of the system design, guide the processing of the signal, towards an optimal rendering of the information in the signal?”. In current recognition systems the complete processing of the signal is dictated by the design of the system. This entails that all possible input has to be considered by the designer of the system, which is impossible in open environments. Instead the system should be able to estimate if and to what extend and how the incoming signal should be processed.

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

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

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

7. correlate with perceptual descriptions of real-world recordings.

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

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

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

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

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

Performance on the datasets is measured with the F-measure on frames. This measure is the harmonic mean between recall and precision, and punishes both failure to include frames for a specific class as well as including too many frames. These measures are chosen based on similarities of the identification task with information retrieval. Because the ambiguity in the start and stop times, the overlap of the recognition results with the annotations will not be exact and the F-measure will be lower due to this mismatch, while
the recognition result is just as valid. For the Amstel station dataset the F-measure is 0.18 for speech-like classes ("speech", "singing", "scream"), partly because of the start and stop time ambiguity, but also because of arbitrary boundaries between the classes. For the "train" and "subway" classes the F-measure is around 0.5. For the Assen dataset the scores are similar, 0.45 for "car", down to 0.02 for "bird", for most classes the precision is high and the recall low. The inter-annotator f-measure for this dataset is 0.46. Though not directly comparable, it indicates that human annotators disagree on the annotations which form the ground truth with the same order as the system (dis)agrees with the ground truth. The signal-driven recognition stage can be complemented with a dynamic network (PhD Thesis M.E. Niessen), which increases the F-measure on average 20% for the Amstel dataset and 33% for the Assen dataset.

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

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

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