

Contents lists available at [ScienceDirect](#)

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Sound event recognition through expectancy-based evaluation of signal-driven hypotheses

J.D. Krijnders*, M.E. Niessen, T.C. Andringa

Artificial Intelligence, University of Groningen, P.O. Box 407, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Article history:
Available online xxxxx

Keywords:

Environmental sound recognition
Spreading activation network
Bottom-up and top-down processing
Audio signal processing
Signal-driven processing
Expectancy-based checking

ABSTRACT

A recognition system for environmental sounds is presented. Signal-driven classification is performed by applying machine-learning techniques on features extracted from a cochleogram. These possibly unreliable classifications are improved by creating expectancies of sound events based on context information.
© 2009 Elsevier B.V. All rights reserved.

1. Introduction

We present the results of an experiment where signal-driven (bottom-up) recognition is combined with knowledge of the context (top-down knowledge) to improve the performance of environmental sound recognition in real-world circumstances. The real-world sonic environment is often referred to as a soundscape, that is, an environment of sounds with emphasis on the way it is perceived and understood by an individual or by a society (Schafer, 1977). Although full soundscape analysis is beyond the scope of this paper, we aim to build a system that can become the basis for an automatic soundscape analysis tool by identifying sound events in real-world environments.

A system that identifies sound events in continuous recordings has additional requirements compared to a system that classifies sound samples, of which is known that they have content. In recognition, a system needs to segment the signal and separate the sources before it can classify them (Shinn-Cunningham, 2008; Grif-fiths and Warren, 2004; Roman and Wang, 2006; Barker et al., 2005).

Furthermore, a system that analyzes soundscapes has to deal with transmission effects such as concurrent sources and reverberation. Reverberation results in a mixing of the target sound with a time delayed version of itself. Therefore, it precludes the successful application of feature vectors that describe the whole spectrum, such as Mel-frequency cepstral coefficients (MFCC's) and the continuous wavelet transform (CWT). MFCC's have been shown to be

very successful for single-source, non-reverberant speech recognition (O'Shaughnessy, 2008). Moreover, MFCC's and CWT have been used successfully in environmental sound recognition provided that the recordings contain a single, clean source (Cowling and Sitte, 2003). However, this is an unrealistic approximation for actual environmental sounds.

Real-world environments pose another problem on techniques used in speech recognition. Speech recognition relies on a strong temporal ordering, but for environmental sounds this ordering is far weaker. Speech recognition techniques exploit this ordering by applying hidden Markov models to find the best model sequence (O'Shaughnessy, 2008). In the case of non-speech sound recognition, such as music genre determination, it has been shown that temporal information is not necessary to recognize genre (Aucouturier et al., 2007). However, music genre determination does not require the detection of sound events and is therefore not suitable to describe the sonic environment in detail.

Another method for sound analysis, the bag-of-frames (BOF) method, has been shown to be able to identify scenes from real-world recordings (Aucouturier et al., 2007). However, the BOF method is not designed to represent details about individual sources 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 classification of sound events.

In contrast to the BOF method and whole spectrum descriptors, the methods we present in this paper segment the spectrum on the basis of the local spectro-temporal properties. Segments are likely to stem from a single source when they are based on local properties. The robustness and reliability of these segments, called signal components, are improved with grouping principles from auditory

* Corresponding author. Tel.: +31 50 363 6955; fax: +31 50 363 6687.

E-mail addresses: j.d.krijnders@ai.rug.nl (J.D. Krijnders), m.niessen@ai.rug.nl (M.E. Niessen), t.andringa@ai.rug.nl (T.C. Andringa).

scene analysis, such as common onset, common offset and common frequency development (Bregman, 1990; Ellis, 1999). These groups are classified as sound events using a naive Bayes classifier.

Systems that perform environmental sound recognition, with similar preprocessing as proposed in this paper, are applied commercially in real-life situations (van Hengel and Andringa, 2007). These systems extract one bit of information from their environment, namely: “is there verbal aggression, or not?”. The more general problem of environmental sound recognition is more complex, but shares some properties with information retrieval, especially with associative retrieval (Crestani, 1997). For both applications it is desirable to retrieve relevant information that is associated with some information item, such as a user query. In environmental sound recognition, retrieval corresponds to estimating the presence of sources and processes from the signal's history and its environmental context. Similar to information retrieval, it is not essential to recognize all sound sources (documents). Instead, it is important to determine sufficient information about the environment to extract relevant parts of the signal, that is, being able to answer the question that spawned the search. Because of the similarities between environmental sound recognition and associative information retrieval, we use the same measures of success, such as precision, recall, and the F -measure.

The data set used in this paper is created to test aggression detection systems. However, the content is fairly rich, since it is recorded on a busy train station. Therefore, it includes problems of real-world environments, such as transmission effects and ambiguous sound events. For example, the sound of a train and a subway are very similar. Based on the sound alone, even human listeners have problems identifying the event correctly, unless they are provided with context (Ballas and Howard, 1987). An automatic system that identifies sound events in real-world situations can benefit from contextual information to recognize events, similar to humans listeners.

To approach this human strategy, we propose a method inspired by cognitive research (Quillian, 1968; McClelland and Rumelhart, 1981). This method constructs a dynamic network that keeps track of both bottom-up signal information and contextual knowledge. By using more information than what can be known from the signal at each point in time, the system is not only more robust to noise, but it can also distinguish between sound events that are similar in acoustic structure but different in meaning (Niessen et al., 2008). The nodes of the dynamic network represent information about sound events at different levels of complexity. Whenever new signal-driven information becomes available, the information in the network is updated. Subsequently, this information is used to form expectancies of future sound events.

The paper is divided in five sections. The following section discusses the data set. Furthermore, we explain the signal-driven processing using filter banks, signal components and machine learning. The third section describes how contextual knowledge is learned and incorporated in the system. Section 4 discusses the results of the signal-driven and the combined system, which uses knowledge of the context on top of the signal-driven information. Finally, in the fifth section we explain and discuss the results and give suggestions for future work.

2. Signal-driven processing

2.1. Data set

The data set (Zajdel et al., 2007) consists of 40 enacted scenes from 16 different scenarios, which last between 1 and 2 min each. The total duration of the recordings is 54 min. The scenes were acted by professional actors (three men, one woman) on a platform

of the station Amsterdam Amstel. The recordings are distorted by reverberation, because the Amstel station is a glass box. 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. For example, the ‘pickpocket’ scenario was played out twice with different actors. All scenarios were played out twice or more. The 16 scenarios were based on stories occurring at stations, such as friends meeting, enthusiastic football supporters and diverse forms of verbal aggression and vandalism. The scenes were recorded by 8 microphones (16 bits, 44.1 kHz sampling rate), of which one was used for this study. This microphone was located about 2 m from the centre of the action and about two meters from the subway track. Saturation of the microphones was checked not to occur when goods trains passed. The scenes were also captured by three calibrated cameras.

The 40 scenes were annotated by the authors for seven classes (see Table 1), based on audio and video. The start and stop times of each event were annotated. For subways and trains, and for some speech, singing and screams, these times were ambiguous, because it is hard to indicate the exact time these events become loud enough to be detectable. The assignment of classes included subjective decisions like whether or not a sound is speech or a scream. These decisions were left to the annotator. Therefore, the annotations are far from perfect.

2.2. Cochleogram

To analyze the sound signal, we convert the time signal to a time–frequency representation. A gammachirp filter bank (Irino and Patterson, 1997) performs this conversion:

$$h_{gc} = at^{N-1} e^{-2\pi b B(f_c)t} e^{j(2\pi f_c t + c \log(t))} \quad (1)$$

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

$$B(f_c) = 24.7 + 0.108f_c \quad (2)$$

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 ms. The energy is compressed logarithmically and expressed in decibel (dB). We call this representation a cochleogram.

2.3. Tone-fit and pulse-fit

After converting the time signal to the cochleogram domain, tones and pulses are extracted from the cochleogram. We apply channel dependent matched filters that respond to ideal tones and pulses. The derivation of these filters is depicted in Fig. 1. For each channel an ideal sinusoid is generated and processed using the filterbank (Fig. 1a). Subsequently the width of the response in frequency at a threshold under the energy maximum is

Table 1

The annotated classes and the number of their occurrences in the data set.

| Class | # |
|------------------|-----|
| Singing | 82 |
| Speech | 521 |
| Train | 15 |
| SubwayDoorSignal | 14 |
| Subway | 40 |
| Kick | 26 |
| Scream | 290 |

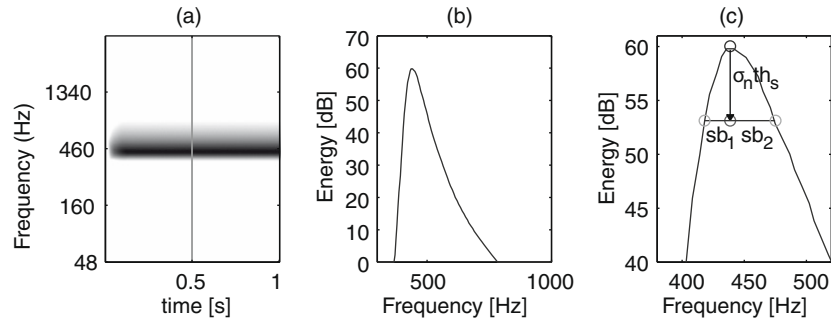


Fig. 1. Calculation of TF filters. (a) Cochleogram of ideal tone. (b) Cross-section at $t = 0.5s$. (c) Detail of (b) with filter parameters.

calculated (Fig. 1c). This threshold is set to twice the standard deviation of the log energy of white noise in a channel. 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. Application (Fig. 2) of the filters is the complementary process, the energies at the widths below (before) and above (after) the time–frequency point are averaged. The difference between the energy at the point for that channel and the average forms the filter output. The application of the filters to the cochleogram results in two representations. They reflect to what extent 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 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 Fig. 3).

2.4. Harmonic complexes

If possible, the tonal signal components are combined into harmonic complexes (HCS) by selecting more and more tonal signal components that comply with the properties of a harmonic complex. 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 components. 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 :

$$S = n_{sc} + b_{f_0} + n_h - \sum_{sc} \text{rms}_{sc} - \sum_{sc} \Delta f_{sc} \quad (3)$$

where n_{sc} is the number of signal components in the group, b_{f_0} is one or zero depending on 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, listed in Table 2. These features will be used in the signal-driven recognition stage.

2.5. Broadband events

Evidence for broadband events, such as trains, is determined by an algorithm that searches for slow broadband changes in the signal. These events have to satisfy a combination of criteria. The change in signal must last at least 2 s, 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 the time. This level of 95% assumes that each channel is dominated by background noise at least 5% of the time. The criterion is fairly safe and works well in practice, assuming that a temporal scope can be chosen appropriately. We chose a temporal scope that was as long as the whole file (about a minute). The energy must exceed the background by three standard deviations of white noise in that channel.

The events that comply to the aforementioned criteria 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 five remaining features are the first five cepstral coefficients that describe the spectral envelope. The three properties for the five bands are only com-

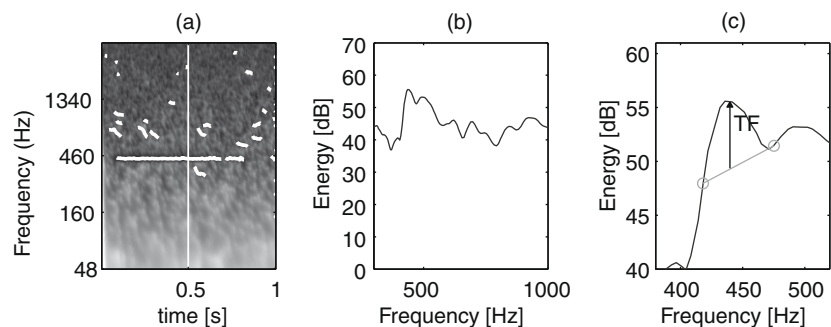


Fig. 2. Application of TF filters. (a) Cochleogram of ideal tone in zero dB local SNR white noise. (b) Cross-section at $t = 0.5 s$. (c) Detail of (b) with filter parameters.

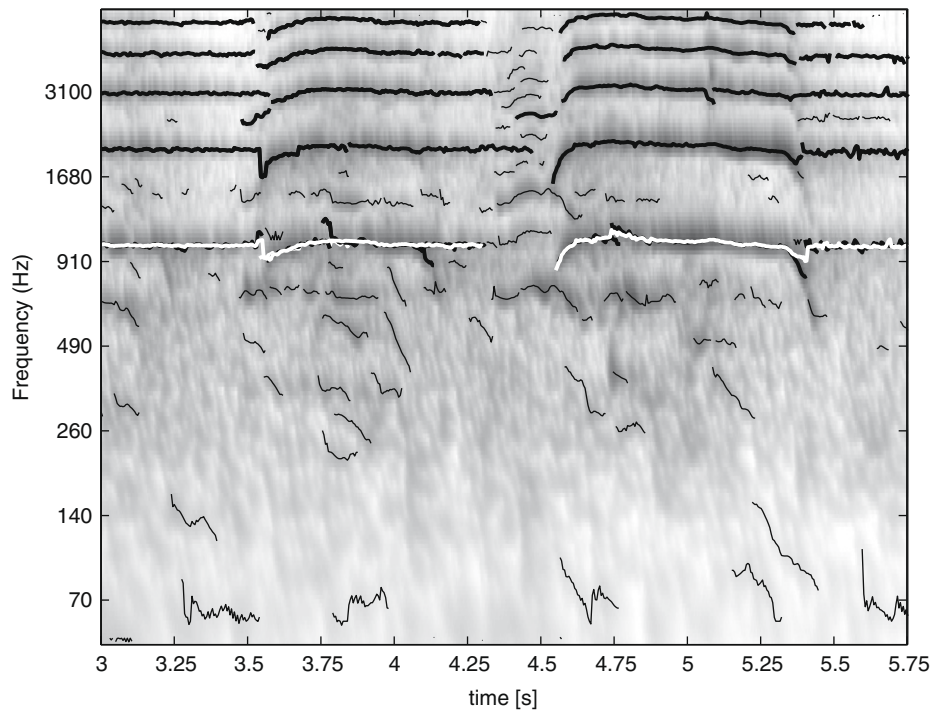


Fig. 3. The background shows the cochleogram of a few screams and a departing subway train. The black lines indicate signal components, the thick black lines are grouped together to form harmonic complexes, and the white lines indicate their fundamental frequencies. Spurious contributions, due to pattern in noise are inevitable, but they can be discarded if they do not contribute to patterns at higher levels of aggregation.

Table 2

The features extracted from each harmonic complex. Features 1, 2, 3 and 9 are picked by the authors to indicate the strength of the harmonic complex. The other features are selected from van Hengel and Andringa (2007), Zajdel et al. (2007) to discriminate speech, scream, singing and subwayDoorSignal.

| | |
|---|---------------------------------------------|
| 1 | Length in seconds |
| 2 | Score from Eq. (3) |
| 3 | Number of signal components |
| 4 | Mean energy under the signal components |
| 5 | Std deviation of energy under the signal |
| 6 | Spectral tilt of the signal components |
| 7 | Mean fundamental frequency |
| 8 | Standard deviation of fundamental frequency |
| 9 | Feature 2 divided by feature 1 |

puted 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, such as speech. 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 distribution can be different between subway trains and normal trains. The third property is the average foreground-to-background ratio for each band, which reflects the total energy per band compared to the background. This property might differentiate between nearby and far-away events.

3. Dynamic network model

The signal-driven processing provides hypotheses based on information in the signal. However, real-world sound recordings, such as in the data set used in this study (see Section 2.1), are distorted by transmission effects similar to broadband noise. Furthermore, some sound events can produce similar acoustic signals, but

have a different meaning. For example, although speech and screams result in a similar acoustic pattern, they differ in meaning, and require a different response. Distortions due to transmission effects and ambiguous sounds might lead to erroneous hypotheses, because the signal provides too little information to allow a correct inference. Knowledge about the environment and the context of a sound event can be used to improve the classification through predictions. Specifically, past sound events can lead to expectancies of the sound events that will follow. If a signal-driven hypothesis matches an expectancy, it is more likely to be correct. In this section we present a model that creates expectancies of sound events and evaluates the signal-driven hypotheses based on these expectancies. The description of the way the model operates is given in more detail in Niessen et al. (2008).

3.1. Knowledge network

The knowledge about the environment is learned in a supervised training phase and stored in a static network, referred to as the knowledge network. This knowledge network is similar to semantic networks used in information retrieval (e.g. Crestani, 1997; Van Maanen et al., in press). Information retrieval is concerned with retrieving relevant information associated with some information item, such as a user query. Therefore, semantic relations, like similarity, between pieces of information are stored in a semantic network. Nodes in this network represent information items, and the connections between the nodes represent the relations between these pieces of information. In automatic sound recognition, a node could represent a speech event, or a whistle followed by a train arrival. Furthermore, the relation between events are represented by the strength of their connection.

Annotations of sound recordings (see Section 2.1) are used in the supervised training phase to learn relations between sound events. When two sound events occur within a certain interval, they are combined in a separate node. The relation between the

node that represents the sequence of the events and the nodes that represent the individual 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 with the inverse frequency it occurs in other documents. 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 A is encountered often in combination with some sound event B , and little with other sound events, it is important in the event sequence $S : A - B$. Accordingly, the strength between the sound event A and the event sequence S is:

$$w_{A,S} = \text{tf} \cdot \log \left(\frac{N}{n} \right) \quad (4)$$

where N is the total number of sequences, n is the number of sequences in which A occurs, and the term frequency is given by:

$$\text{tf} = \frac{f_{A,S}}{\sqrt{f_A}} \quad (5)$$

where $f_{A,S}$ is the number of occurrences of A in S , and f_A is the total number of occurrences of A in the training set.

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. However, for some sequences the order can be very indicative. For instance, in the data set there are trains departing, which are always preceded by a whistle of the conductor. Hence, if a whistle is heard, a strong expectancy of a train departing should arise. To capture the expectancies of fixed sequences, we determine whether the sound events that constitute a sequence have a strong bias to a specific order. For these fixed sequences the mean time difference between the events is used in a function to calculate the expected value of the second event in the sequence. In other words, the first sound event of a fixed sequence primes the network for the second sound event after a learned time interval. In the next subsection, we will show how this expected value is computed for both ordered and non-ordered sequences.

3.2. Dynamic network of hypotheses

Once the knowledge network is fully trained, it is used in the operation phase to evaluate signal-driven hypotheses of sound events. Each signal-driven hypothesis is initiated as a node in the dynamic network. The dynamic network has three levels of representation. The hypotheses at the first level represent detected structures in the signal, as described in Section 2. The second level consists of hypotheses of possible sound events that explain the structures. Finally, the third level contains hypotheses of sequences of events, as described in the previous subsection. Fig. 4 shows an example of a network with two signal-driven hypotheses about structures in the signal, their connections to possible sound events that caused them, and a sequence of which they might be part.

When a new signal-driven hypothesis is added to the dynamic network, the configuration of the network is updated. First, the hypothesis that represents a structure in the signal is connected to hypotheses of sound events that can explain the structure. The strength of this connection is determined through naive Bayes classification of the structures, as will be described in Section 4.1. Next, the connections of these sound events to possible event sequences are retrieved from the knowledge network and added to the dynamic network. The connections in the network are only between hypotheses at different levels, as can be seen in Fig. 4. As a

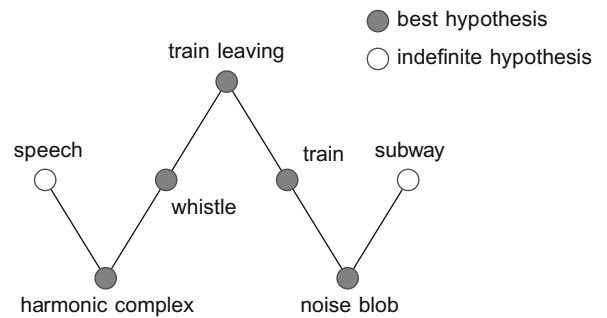


Fig. 4. An example of a network with two signal-driven hypotheses about structures in the signal. Both hypotheses are connected to two hypotheses of sound events that can explain the structure. Two of these hypotheses are part of an event sequence, increasing the support for the sound events that are part of the sequence.

consequence, the dynamics and hierarchy of the network are captured by the hypotheses and their connections.

3.3. Activation

The activation value of a hypothesis is a weighted sum of its input activation from connected hypotheses. The activation of a signal-driven hypothesis is spread through the network after the configuration is updated. As a result, every hypothesis in the network holds a confidence value after spreading the activation. A description of the details of the spreading activation algorithm can be found in Niessen et al. (2008). The activation values of all hypotheses in the network decrease with time when they get no reinforcement from signal-driven evidence.

The activation values of event sequences are used to compute the expected activation of events that are not active yet, and are part of the sequence. For example, in a non-fixed event sequence such as singing and speech, of which speech is already identified, the expected activation of a singing event is calculated by multiplying the activation value of the event sequence with the connection strength between the sequence and the type of event (see Formula (4)). Since the activation value decays with time, the expected value is smaller when the other event of the sequence occurred longer ago.

For fixed event sequences, the expected value will furthermore be dependent on the time when the event is expected:

$$\hat{A}_i(t) = w_{ij} A_j(t - \Delta t) e^{-\frac{(\Delta t - \bar{T})^2}{2\sigma^2}}, \quad (6)$$

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

4. Experiments

To test the system we apply it to the data set of 40 realistic recordings (see Section 2.1). In the first experiment only the signal-driven classification is used. In the second experiment these results are used in the expectancy-based dynamic network.

4.1. Experimental setup

All 40 audio files were processed with the methods explained in Section 2 to extract harmonic complexes and their features (see Table 2). The harmonic complex with the highest score and overlap was selected for each annotation and labeled according to the

annotation. Harmonic complexes that do not overlap in time with an annotation were labeled as noise. Harmonic complexes that do overlap with an annotation, but do not have the highest score, are discarded. From these files, 40 pair files were generated, of which 40 files were used for training, all with the instances from one scene left out, and 40 files were used for testing, with instances from the scene that was left out, thus creating a leave-one-scene-out set.

Because of the strong link with information retrieval (see Section 3.1) we use performance measures from that field, such as precision and recall, to quantify the performance of our system. Precision is a measure for the fraction of time our detections were correct, and recall is a measure for the fraction of detections we should have made are actually made. The F -measure is the harmonic mean of these two, giving a single performance measure. The formula's are given as:

$$\text{precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

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

For the first experiment a naive Bayes classifier from the Weka toolbox (Witten and Frank, 2005) was trained on the leave-one-scene-out training file and tested on the corresponding testing file. The labeling and classification of the noise regions was performed in the same way as the harmonic complex classification. The results of both classifications were taken together to create a single result set.

In the second experiment the supervised training of the knowledge network (see Section 3.1) was performed on the same data as the classifier, that is, the annotations of the leave-one-scene-out training file. Hence, the test set was not used for training. On average 18 different types of sequences were encountered in the train-

ing set. These sequences are composed of the seven classes listed in Table 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 is very large, ranging from 2 to 730. For the testing, the results of the classifier were input for the dynamic network of hypotheses (see Section 3.2).

4.2. Signal-driven results

The white bars of Fig. 6 show the F -measure, the precision, and the recall of the signal-driven classification. The overall F -measure is 0.37, the overall precision is 0.39, and the overall recall 0.34. The results of one of the scenes are shown in the lower panel of Fig. 5.

Part of the errors arise from alignment errors of the annotations. For example, all detections of the subway trains are longer than the annotations. This problem is hard to solve, because the annotators did not agree when on the moment when a train is first and last detectable. Therefore, the detection cannot agree with both annotators. A partial solution would be to introduce “don't care” regions around annotations where the algorithm is not punished for incorrect detections.

The major groups of confusion are between trains and subway trains, and between speech, singing, and screams. These confusions may partially be caused by confusion in the annotations. The distinction between a train and a subway train is hard to make based on audio recordings, even for a human annotator. The boundaries between the classes speech, singing and scream are fairly arbitrary, which causes confusion in the annotations.

The F -measure on the kick class is small because it is neither a harmonic nor a broadband sound. The features we have used were not suited for describing these pulse-like sounds.

The systems calculations run at about real-time on a modern PC (2 GHz dual-core). However, the current Matlab code is not optimized. Based on similar systems optimized for speed (van Hengel and Andringa, 2007) we estimate that the performance could be around four times real-time on the same machine.

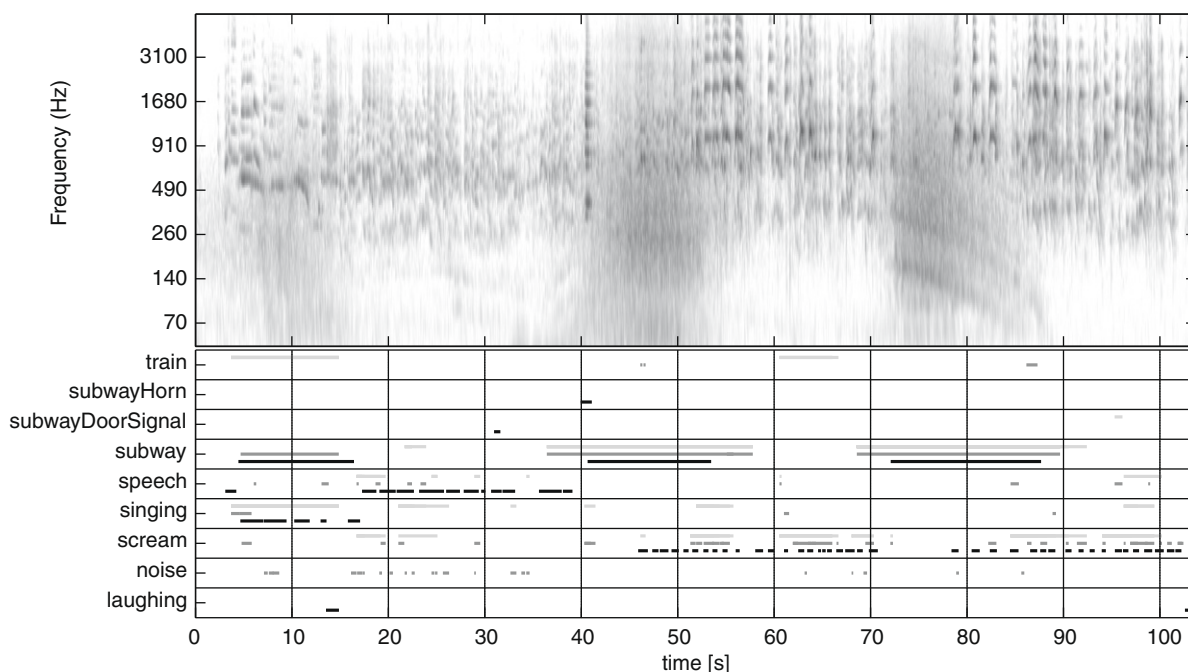


Fig. 5. The upper panel shows the cochleogram of a complete scenario. A darker color corresponds to more energy. In the first 40 s there is some speech and singing. At $t = 41$ s a subway horn occurs, which is followed by the noise event of a subway train passing by. Around $t = 55$ s four clear screams occur, followed by a few more muffled ones. At $t = 72$ s a subway train enters the station, again followed by screams. The lower panel shows the annotations and detections for the different classes. The lower, black, lines represent the annotations, the middle, gray, lines the signal-driven detections, and the upper, light-gray, lines the final, expectancy-based results.

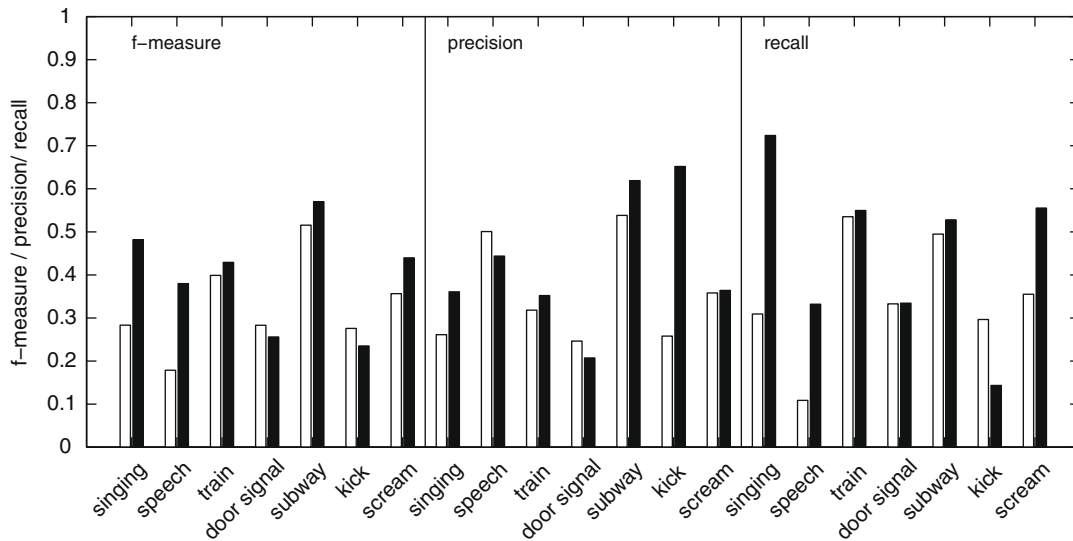


Fig. 6. The results of the signal-driven classification (white bars) and of the expectancy-based results (black bars).

4.3. Expectancy-based results

The black bars of Fig. 6 show the performance measures for the classification including the dynamic network. The overall F -measure improves to 0.45 (20%), the overall precision to 0.42 (8%) and the overall recall to 0.49 (44%). The main improvement is in the recall of the classes that have more harmonic content (singing, screams and speech), because events of these classes are more likely to be of the same class as their neighbors. As a consequence, 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 class and the precision of the speech class increase. However, the increase in precision is moderated by other erroneous changes. As a result, the overall precision does not increase substantially. Due to the ambiguous nature of some of the classes and the acoustic environment a high F -measure is not achieved. So we conclude that the inclusion of the dynamic network leads to a result more consistent with manual annotation.

5. Discussion

In the previous section we have demonstrated that the combination of signal-driven algorithms and a dynamic network of hypotheses results in a recognition improvement for most sound event classes compared to an exclusively signal-driven method. Especially the classes that have similar signal structures, and hence rely more on context for their interpretation (screams, speech and singing), are better identified in the combined approach. Classes that are already identified well by the signal-driven algorithm (subway and train) gain little improvement from the dynamic network. Finally, both classes that occur infrequently, and hence have little training examples, and classes that are not yet captured well by the signal features, show a small performance reduction.

We have shown that the use of a dynamic network model improves the overall performance of environmental sound recognition. However, apart from sound event recognition, this model provides more diverse ways to analyze a soundscape. More specifically, through hierarchical relations in the network, recognition of sound events can lead to abstract descriptions of the soundscape. This introduces the possibility to describe complex activities in the neighborhood of the microphone with complex and efficient linguistic descriptions (Guastavino, 2007).

Furthermore, the input information that is presented to the network is not limited to a specific modality. Niessen et al. (2009) show that the dynamic network model can also be used to improve visual robot localization. Because the model can receive input from different modalities, it can combine multiple modalities in a single system. For example, if input from one modality, such as images, is insufficient, input from other modalities, such as audio or GPS, can help to generate predictions. In future work we plan to integrate information from multiple sources of knowledge to reach more reliable event recognition with richer descriptions.

One of the major problems in the development of environmental sound recognition systems that operate in real-life situations is the lack of large, diverse, and annotated data sets that can be used for training and testing. This is one of the reasons that we tested on a data set that represented only a single location and a limited amount of events. The main problem of constructing more realistic data sets is the large number of different events that can occur outdoors and the associated time it takes to annotate a representative set. The development of an annotation tool for soundscape research is helpful in this respect.

Another problem in environmental sound recognition is performance evaluation. We have used the measures precision and recall to quantify the performance, since these measures are common in the related task of information retrieval. We calculated these measures in terms of the temporal overlap of annotations and classifications. However, if we were to apply these measures in line with the field they were originally developed for, we should only check whether or not an annotated event was detected. We have chosen for overlap instead of presence, because the combination of the short annotations of speech events in combination with small temporal alignment errors made the attribution difficult. Allowing some flexibility in matching system detections with hand annotations may alleviate this problem. This however requires a more formal justification, before it can be applied.

The current system shows that it is possible to build a recognition system that captures many of the events of a realistic and minimally constrained sonic environment. The background was completely uncontrolled while the foreground consisted of actors who improvised a range of both social and aggressive activities. We have shown that it is beneficial to use the history of identified sound events to form a context in which the current sonic evidence is weighted. This is done by forming a dynamic network that mimics short-term memory dynamics. The interplay of knowledge-

driven and signal-driven processing is characteristic for human perception. Since human perception is effectual in a wide range of acoustic environments, we consider this interplay a promising approach for robust automatic sound recognition.

Acknowledgments

J. D. Krijnders' work is supported by The Netherlands Organization for Scientific Research under Grant 634.000.432 within the ToKeN2000 program. M. E. Niessen's work is supported by SenterNovem (Dutch Companion Project Grant No. IS053013). This research is also supported by Foundation INCAS³ (Assen, The Netherlands).

References

- Aucouturier, J.-J., Defréville, B., Pachet, F., 2007. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes *J. Acoust. Soc. Amer.* 122 (2), 881–891.
- Ballas, J.A., Howard, J.H., 1987. Interpreting the language of environmental sounds. *Environ. Behav.* 19 (1), 91–114.
- Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. *Speech Comm.* 45 (1), 5–25.
- Bregman, A., 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.
- Cowling, M., Sitte, R., 2003. Comparison of techniques for environmental sound recognition. *Pattern Recognition Lett.* 24 (15), 2895–2907.
- Crestani, F., 1997. Application of spreading activation techniques in information retrieval. *Artif. Intell. Rev.* 11 (6), 453–482.
- 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 Comm.* 27 (3–4), 281–298.
- Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? *Nature Rev. Neurosci.* 5, 887–892.
- Guastavino, C., 2007. Categorization of environmental sounds. *Can. J. Exp. Psychol.* 61 (1), 54–63.
- Irino, T., Patterson, R.D., 1997. A time-domain, level-dependent auditory filter: The gammachirp. *J. Acoust. Soc. Amer.* 101 (1), 412–419.
- McClelland, J.L., Rumelhart, D.E., 1981. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* 88 (5), 375–407.
- Moore, B.C.J., Glasberg, B.R., 1996. A revision of Zwicker's loudness model. *Acta Acust. United Acust.* 82 (2), 335–345.
- Niessen, M.E., Kootstra, G., De Jong, S., Andringa, T.C., 2009. Expectancy-based robot localization through context evaluation. In: *Proceedings of the ICAI 2009, Las Vegas*, pp. 371–377.
- Niessen, M.E., Van Maanen, L., Andringa, T.C., 2008. Disambiguating sound through context. *Internat. J. Semantic Comput.* 2 (3), 327–341. doi:10.1142/S1793351X08000506.
- O'Shaughnessy, D., 2008. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition* 41 (10), 2965–2979.
- Quillian, M.R., 1968. Semantic memory. In: Minsky, M. (Ed.), *Semantic Information Processing*. MIT Press, Cambridge, MA, pp. 216–270.
- Roman, N., Wang, D., 2006. Pitch-based monaural segregation of reverberant speech. *J. Acoust. Soc. Amer.* 120 (1), 458–469.
- Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manage.* 24 (5), 513–523.
- Schafer, R.M., 1977. *The Soundscape, Our Sonic Environment and the Tuning of the World*. Destiny Books, Rochester, VT.
- Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. *Trends Cognit. Sci.* 12 (5), 182–186.
- van Hengel, P., Andringa, T.C., 2007. Verbal aggression detection in complex social environments. In: *Proceedings of AVSS 2007, London*, pp. 15–20.
- Van Maanen, L., Van Rijn, H., Van Grootel, M., Kemna, S., Klomp, M., Scholtens, E., in press. Personal publication assistant: Abstract recommendation by a cognitive model. *Cognit. Syst. Res.* doi:10.1016/j.cogsys.2008.08.002.
- Witten, I.H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, second ed. Morgan Kaufmann.
- Zajdel, W., Krijnders, J.D., Andringa, T.C., Gavrila, D., 2007. Cassandra: Audio–video sensor fusion for aggression detection. In: *Proceedings of AVSS 2007, London*, pp. 200–205.