2

CHALLENGES OF SOUND RECOGNITION IN THE REAL WORLD

Systems that operate in a real-world environment have to process ambiguous and noisy input. Current techniques for sound recognition are mostly designed for specific applications, such as automatic speech recognition. Therefore, they can rely on assumptions about the audio signal, such as the signal being undistorted and of a known type. However, these assumptions cannot usually be met in a real-world environment. Therefore, real-world sound event recognition requires methods to segregate individual sound events from a globally sounding environment. Furthermore, a system that operates in an uncontrolled environment needs to handle transmission effects. To be able to function reliably, the system should be able to adapt to a variety of situations. However, it is not necessary to solve the problems of real-world environments only with signal-driven methods.
2.1 INTRODUCTION

Systems that operate in a real-world environment are confronted with additional challenges compared to systems that operate in a simulated or controlled environment. They have to process an abundance of information, of which not everything is necessarily relevant (Van Hengel and Andringa, 2007). Moreover, sensory information is likely to be ambiguous or noisy. In section 2.2 we evaluate whether classification methods used in automatic sound recognition are suitable to recognize sound events in real-world environments. In section 2.3 we discuss the problem of separating sound events from the background in order to recognize them. The recognition of sound events in real-world environments is further complicated by transmission effects, such as reverberation and concurrent sources. Possible approaches to deal with transmission effects are discussed in section 2.4. Finally, in section 2.5, we conclude with the implications of these challenges for robust real-world sound event recognition.

2.2 STATE OF THE ART IN AUTOMATIC SOUND RECOGNITION

Automatic sound recognition has important (future) applications in fields as diverse as environmental noise monitoring, robotics, security systems, content-based indexing of multimedia files, and human-machine interfaces. Most sound recognition research is aimed at improving one of these application domains, such as speech recognition or music genre classification. Typically, the techniques used in these applications classify a sound sample as one class of a closed set of learned classes, of which the descriptors match the descriptors of the sample best. Because these techniques are applied to a known type of sound, they can apply specialized features to describe the sound (Davis and Mermelstein, 1980). For example, within music genre classification (Tzanetakis and Cook, 2002; Aucouturier and Pachet, 2003) and speech recognition (O’Shaughnessy, 2008), spectral-based features such as Mel frequency cepstral coefficients (MFCCs) capture important information of harmonic sounds, but are not very robust to noise (O’Shaughnessy, 2008). In addition, methods for automatic speech recognition rely on a temporal ordering of the signal, which is exploited by searching for the most probable sequence of hidden Markov models (HMMs, Juang and Rabiner, 1991).
Cowling and Sitte (2003) tested a selection of feature extraction techniques and classification methods for speech and music signals on isolated environmental sounds. All methods perform better on speech sounds than on environmental sound events, because environmental sounds exhibit more diverse acoustic properties than speech. However, the best results (70% classification rate for eight classes) suggest that some classification techniques can be effectively applied to the recognition of isolated environmental sounds. Similar to applications in speech and music, the input is controlled so that the methods can classify it. In other words, the input has content that belongs to a single class that is a member of a limited set of known classes, although the type of sound is different from and more diverse than speech and music. In contrast, Defréville et al. (2006) applied multiple features to classify samples from continuous real-world recordings. Their results varied from 72% to 99% classification rate per class for six classes.

Another method for sound analysis, the bag-of-frames (BOF) method, has been shown to be able to identify auditory scenes from real-world recordings, such as the street where a recording has been made (Aucouturier et al., 2007). However, the BOF method is not designed to represent details about individual events 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 recognition of sound events. For example, the estimated location of a recording, such as a street or a park, can be useful to infer probable sound events that may occur.

Although the combination of whole-spectrum feature extraction and classification has proven useful in problems with a single known signal type, and even in environmental sound recognition, these methods have some attributes that make them less suitable for automatic sound event recognition in real-world environments. Sound classification methods either classify pre-selected samples with one type of sound, or segment a signal into different intervals based on its acoustic properties. However, these intervals do not necessarily correspond to events. A system for real-world sound event recognition needs to segregate the individual events from the background before it can classify the events. Sound events in a real-world environment often co-occur, at other moments no (interesting or recognizable) events may take place. In other words, we do not want to know of certain intervals what type of sound they are, but which events are present at what time.
in a continuous audio signal. In variable real-world environments, a method for sound event recognition cannot assume known and uniform input, as the methods used on speech and music can.

### 2.3 Sound Event Segregation

To be able to recognize individual sound events in a globally sounding environment (see Figure 1.3), the *audio components* that constitute an event need to be separated and grouped from the background. We refer to the joint process of separating components and grouping them as segregation. Practical applications of sound recognition, such as automatic speech recognition, have advanced research in sound segregation methods. Strategies for segregation include spatial separation of audio input from microphone arrays (beamforming, Brandstein and Ward, 2001), and independent component analysis (ICA, Jutten and Herault, 1991). Although these methods are successful in the applications they are designed for, they rely on assumptions that cannot be met in all situations. For example, beamforming approaches are challenged by moving targets and by multiple targets close to each other. ICA assumes that the sound events in a mixture are statistically independent. Consequently, these methods are not suitable as a general approach to sound event segregation. To address the issues of application-driven methods, a field of research has emerged that is inspired by human perceptual mechanisms, called computational auditory scene analysis (CASA, Wang and Brown, 2006).

The common aim of CASA studies is to infer properties of individual sound sources from an auditory scene based on either a mono or a stereo recording. CASA research is primarily aimed at analyzing speech and music sounds.¹ Therefore, the target sound is defined by the problem, namely segregating speech or music sounds, either from each other or from background sounds. However, two important aspects of human auditory scene analysis (ASA) are underrated by these approaches. First, human ASA is not limited to segregating harmonic sounds. In real-world environments people are confronted with many other sound classes as well, such as machine, traffic, and nature sounds. Second, whether a sound is tar-

---

¹ The topic query “computational auditory scene analysis” gives 222 results in ISI Web of Knowledge and about 1850 in Google Scholar. Excluding the results that also contain one or more of the terms “speech”, “music”, “voice”, “harmonic”, and “pitch” leaves only 38 and 86 results respectively.
get or background sound does not depend on the audio signal. Instead, it depends on factors such as the goal and expectancy of the listener (or application): “... all sound sources are potential signals and noises. Whether or not a sound from a particular source is signal (wanted) or noise (unwanted) depends on non-auditory events” (Yost, 1991, p. 16). As a result, a method for general purpose sound event segregation cannot define in advance what the target sound is, and what the background noise. Instead, it should provide hypotheses of what can be inferred from the audio signal based on physical knowledge.

2.4 Transmission effects

In addition to segregating hypotheses about sound events in the audio signal, a system that analyzes sound in a real-world environment has to deal with transmission effects such as concurrent sources and reverberation. Reverberation leads to a mixing of the target sound with a time delayed version of itself. Therefore, both the effects of concurrent sources and reverberation are similar to the problem of sound event segregation, discussed in the previous section. However, the effects of reverberation can be reduced with knowledge of the source and the sonic environment, which facilitates sound event segregation.

The reduction of effects of reverberation have mostly been studied for speech sounds, since the performance of automatic speech recognition (ASR) systems is affected severely by distortions in the signal (Figure 2.1 shows an example of a reverberant speech signal compared to the clean signal). Classifiers in ASR systems, such as hidden Markov models (HMMs), use features that describe the whole spectrum (see section 2.2). When a signal is distorted by a delayed version of itself, irregular frequency dependency patterns of constructive and destructive interference cause rapid energy fluctuations in the frequency content. Hence, the signal descriptors of a reverberant audio signal will deviate from the descriptors in clean training conditions.

One solution to deal with reverberant environments is to reduce the mismatch between the training conditions and the operating conditions. For example, if HMMs are trained on reverberant speech, ASR will perform better in similar reverberant operating conditions than if the HMMs are trained on clean speech (Matassoni et al., 2000). However, the effects of room acoustics vary greatly for different
Figure 2.1: The time-frequency plane (computed with a gammachirp filter bank) of a clean speech signal on the left, and of the same speech in a reverberant environment (reverberation time is 320 milliseconds) on the right. The gray-scale indicates the energy in decibels (dB): darker gray corresponds to more energy. The frequency axis is logarithmic. When harmonic components are stationary, reflections may result in an increased energy of the component, and a longer duration. Reflections of variable components cause irregular distortions in the signal.

environments. Different parameters, such as the size of the room, the material on the floor and walls, and the temperature, influence the acoustic characteristics (Kuttruff, 1979). Therefore, this type of ASR system requires training data that match the characteristics of the operating environment. Couvreur and Couvreur (2004) propose a method where acoustic models are trained on speech under different, simulated reverberant conditions. During operation of the ASR system the model that matches the operating conditions best is selected. They show an improved performance on simulated reverberated speech compared to an ASR system trained on clean speech. However, the improvement on realistic data is not as high as on the data with simulated reverberation, because of the discrepancy between real reverberant and simulated reverberant speech.

Another approach to resolve the discrepancy between the training data and reverberant data is to recover the clean speech from the reverberant signal, instead of
adjusting the training data to the operating conditions. An inverse filter is applied to the reverberant signal to remove the distortion caused by reflections, based on the estimated impulse responses of the environment. However, inverse filtering relies on known and stable acoustic characteristics of the environment. As a consequence, methods based on inverse filtering are not robust to changes in the environment, such as the position of the source or the microphone (Radlovic et al., 2000).

Other methods have been designed that are more robust to reverberation in an unknown, but stable, environment. For example, cepstral mean subtraction (Kinoshita et al., 2009) can handle early reflections in ASR. Additionally, late reflections can be suppressed through spectral subtraction (Wu and Wang, 2006; Kinoshita et al., 2009). When the environmental conditions are sufficiently stable, these methods improve the results of ASR.

To test whether it is possible to assess the reverberation level of a monaural audio signal in an unknown and possibly variable environment, we developed a method to classify the reverberation level of speech signals. Reverberation causes an increase in the variation of the energy and frequency of harmonics in speech. Hence, features that capture this variation can be useful to estimate the reverberation level of a signal without a priori knowledge of the environment. We designed and measured such features on speech samples with different levels of reverberation. Clean speech was artificially reverberated to be able to test a controlled set of conditions.

2.4.1 Methods
A common measure for the level of reverberation is the reverberation time $T_{60}$. The reverberation time is defined as the time for the sound energy level to decay 60 dB after the excitation has ended. We computed nine different levels of reverberation using the Eyring-Norris equation (Eyring, 1930; Norris and Andree, 1929; Kuttruff, 1979):

$$T_{60} = \frac{0.161V}{4mV - S\ln(1 - \bar{a})},$$

where $V$ is the room volume in cubic meters, $m$ is a vector with air absorption coefficients for the frequency bands, $S$ is the total surface area, and $\bar{a}$ is the mean wall absorption coefficient. We assumed a fixed room size of 10 by 12 by 3.5 meters.
and a constant temperature and humidity of 20 °C and 60% respectively. Hence, the mean wall absorption coefficient was the only variable parameter. Values were assigned to this parameter such that we had a collection of nine reverberation levels ranging from no reverberation to a reverberation time of approximately 1.6 seconds.

The reverberation level can also be expressed by the reverberation radius or distance (Kuttruff, 1979). The reverberation radius is the distance from the speaker or microphone to the sound source for which the energy contribution of the direct sound and the reflected energy are equal. A more reverberant environment coincides with a smaller reverberation radius. Naturally, the reverberation radius is strongly correlated with the reverberation time. We also computed the reverberation radius, so the sound samples could be labeled as either clean or reverberant. We regarded clean speech as speech measured inside the reverberation radius and reverberant speech as speech outside the reverberation radius.

The parameter values used in the Eyring-Norris equation were used as input to the shoebox model, which simulates an impulse response in a rectangular room, a shoebox. The shoebox model is an implementation of the image source method of Allen and Berkley (1979). The speaker and the listener or microphone are modeled as two points in space. Apart from the direct sound, specular reflections are computed using mirrored image sources. An impulse response is obtained for every image source. The final impulse response describing the room is computed by combining all individual impulse responses, which are received at different delay times. This impulse response is convolved with the speech signal, resulting in reverberant speech. The speech is processed with a gammachirp filter bank (Irino and Patterson, 1997), which results in a logarithmic time-frequency representation (a cochleogram, see appendix B). Figure 2.1 depicts a cochleogram of a clean speech sample on the left, and a cochleogram of a reverberant speech sample on the right.

We expected that the effect of reverberation on speech can be measured directly in the audio signal. Since we want to test whether we can measure reverberation in an uncontrolled environment, the features used for the classification must have no parameters that require knowledge of the room characteristics. One prominent effect of reverberation on speech is the attenuated salience, that is, the attenuated stability in both the frequency and the energy, of the harmonics (Darwin and Hukin, 2000). Therefore, voiced speech was located based on the selection of harmonic
Table 2.1: Features that indicate the reverberation level on a harmonic track \((h)\). \(E_h(t)\) is the energy development and \(f_h(t)\) is the frequency development of a harmonic track in time. MA is the moving average of an energy or frequency track and P is its polynomial. See appendix C for the calculation of the features.

<table>
<thead>
<tr>
<th>Energy variation</th>
<th>Harmonic energy salience</th>
<th>Harmonic frequency salience</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak rate ((PR))</td>
<td>(\Delta E_h(f))</td>
<td>(\text{Var } f_h/\text{MA})</td>
</tr>
<tr>
<td>(\text{Var } E_h)</td>
<td>(\text{Var } \Delta f_h)</td>
<td>(\text{Var } f_h/P)</td>
</tr>
<tr>
<td>Mean (\Delta f_h)</td>
<td>(\text{Var } f_h/P)</td>
<td>(\text{Var } f_h/P)</td>
</tr>
</tbody>
</table>

complexes—a superposition of co-occurring harmonics—in the cochleogram, using the algorithm presented in Krijnders et al. (2007). Subsequently, the fluctuation in energy and frequency of the first five harmonics of the harmonic complex was measured. These harmonics can be better resolved from the cochleogram because of the logarithmic frequency scale, and are hence more reliable.

We measured the energy and frequency fluctuation through seven features, which are summarized in Table 2.1. The energy variation was measured through the peak rate of the harmonic track \((2.2a)\) and the energy variation of the harmonic compared to its smoothed version \((2.2b)\). Both values are expected to increase at higher reverberation levels. In addition, the energy contributions of echoes cause less distinct harmonics. This effect was captured by calculating the energy slope of the harmonics \((2.2c)\), and the variation \((2.2d)\) and mean \((2.2e)\) of the width of the harmonic compared to an ideal sinusoid. In other words, the energy slope is less steep in reverberant speech, and the harmonic covers a broader frequency range. Reverberation effects can be found in the time-frequency space as well. The short-time development of the harmonic is distorted by echoes, causing a less smooth harmonic track. Therefore, the track variation was measured compared to its smoothed version \((2.2f)\), and to an approximation of a clean harmonic track \((2.2g)\). The calculation of all seven features is worked out in appendix C.

2.4.2 Experiment

Part of the Aurora database (Hirsch and Pearce, 2000) was used to validate the six features. Artificial reverberation was added to 685 randomly selected clean sound
samples with a mean duration of 1.5 seconds, spoken by 214 different speakers, both male and female. The reverberation was computed at nine levels, equivalent to reverberation times distributed roughly linearly between 0 and 1600 milliseconds. As we expected, most of the 685 sound samples showed a significant correlation of at least one feature with the reverberation time. Only 6% did not show a relation for any of the features. However, the predictive strength of the features for individual sound samples is no direct indication for the general classification of the speech samples as either clean or reverberant. To test classification, global thresholds need to be determined in a training set and used to classify a test set.

Since the 685 speech samples were reverberated at nine levels, a total of 6165 samples could be used for classification. After the dismissal of samples in which we could not measure one or more features\(^1\), 5189 samples were left. All samples within the reverberation radius \((T_{60} = 0.22\) seconds\)), the two lowest levels, were labeled as clean, and all samples outside the reverberation radius, the other seven levels, were labeled as reverberant. The data was split in a part for training (33%) and a part for testing (66%). In addition, continuous read speech of six speakers was recorded using a close-talking microphone. This data was split into samples of similar length to the Aurora database, and resampled to an equal sample frequency. The speech samples were artificially reverberated in the same way as the other data. Again, part of the data that was unfit was removed, and 2377 samples with a mean duration of 2 seconds were left. These samples served as an extra test set, which can show the robustness of the features. Finally, both data sets were also labeled with a different threshold for the reverberation time \((T_{60} = 0.7\) seconds\)) to test the performance on a more balanced design.

Numerous methods exist to test the classification accuracy of features. We used a support vector machine (SVM), since it is known to be less prone to problems of overfitting than some other methods (Duda et al., 2001). In training, an optimal separating hyperplane, or threshold boundary, is determined. The support vectors are the speech samples that are closest to the hyperplane, and hence are most difficult to classify. The mapping of the data to a higher-dimensional space is dependent on the type of kernel, that is, the mapping function, which can be defined by the

\(^{1}\) In these samples the harmonic complexes were not sufficiently salient to be segregated by the algorithm.
user, or selected from one of the standards. For our data we use a standard linear kernel. The number of support vectors is an indication of the complexity of the classification. During the testing phase, the speech samples are mapped onto these support vectors. Since the test samples are labeled as well, the classification can be compared to the labels, resulting in a performance measure.

2.4.3 Results

The speech samples from the Aurora database were split randomly into a training set of 1744 samples and a test set of 3445 samples. The seven features (one of which, $\Delta E_b(f)$, has two values) were computed on the first five harmonics, resulting in 40-dimensional data. The skewedness of the first data set—22% of the speech samples was clean and 78% was reverberant, because the reverberation radius corresponds to a relatively low reverberation time of just over 200 milliseconds—is accounted for by using prior probabilities to weight the class error contributions. The SVM was trained on the training samples, resulting in a classifier with 363 support vectors. The rest of the speech samples of the Aurora database was tested on the trained classifier. The performance, or accuracy, of the classifier was 92%. The additional speech samples of our own recordings were also tested on the classifier, with a performance of 87%, only a few percent less. The same procedure was applied to the second, balanced, data set, resulting in an accuracy of 80% on the Aurora samples, and 70% on the recorded speech samples. The results are summarized in Table 2.2. Figure 2.2 gives the detailed results of one of the results (the Aurora data with a labeling threshold of $T_{60} = 0.7$ seconds). The classification of the samples is skewed toward the more reverberant samples, because the features do not develop linearly with the reverberation time. In a separate experiment we tested the predictive strength of the features, and found that they overestimate the reverberation time below approximately $T_{60} = 0.8$ seconds, while the differences between higher reverberation times cannot be predicted (see appendix C). This effect is explained by the difference in the effect of reverberation on the audio signal, which is greater for smaller reverberation levels.

Different classification methods could be chosen for this problem, or different settings for the SVM. For example, if the size of the training set is increased, the performance on the other Aurora speech samples increases, but the performance
Table 2.2: Results of the SVM classifier on two data sets (Aurora and continuous speech) with two designs (unbalanced for the reverberation radius $R_{\text{reverb}}$ and balanced).

<table>
<thead>
<tr>
<th>$T_{60}$ threshold</th>
<th>Accuracy Aurora data</th>
<th>Accuracy recorded data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.22 sec. ($R_{\text{reverb}}$)</td>
<td>92%</td>
<td>87%</td>
</tr>
<tr>
<td>0.7 sec.</td>
<td>80%</td>
<td>70%</td>
</tr>
</tbody>
</table>

on the extra test set decreases. We are not interested in optimizing the classifier on a particular data set, but in the separability of any reverberant speech using the features. Hence, the classification with an SVM using a linear kernel gives an indicative performance result.

2.4.4 Discussion

Although many methods to assess or resolve the effects of reverberation are successful in improving ASR, they generally cannot be used for different applications. Since these methods are designed for ASR, they utilize the common assumptions of ASR that cannot necessarily be met in an uncontrolled environment. First, methods that adjust the training conditions to reverberant conditions or apply inverse filtering cannot deal with variable or unknown conditions. For example, the model of Couvreur and Couvreur (2004) is not tested on speech signals that are affected by transmission effects other than reverberation, such as background noise or concurrent sources. Second, blind dereverberation methods based on spectral subtraction and blind reverberation classification methods like those presented in this section rely on the presence of speech to estimate the reverberation components (Wu and Wang, 2006; Kinoshita et al., 2009). However, if a system must recognize sound events in a continuous audio signal, no assumptions about the presence of a specific sound event can be made. Therefore, dereverberation methods should be extended to include estimation of reverberation for impact sounds and broadband sounds, not only for tonal sounds.

In general, if researchers want to improve sound event recognition in unknown conditions, they should focus on developing robust techniques for online blind dereverberation of mono-signal input with unrestricted content. More specifically, the common experimental paradigm for resolving distortions should extend to
Figure 2.2: Detailed results of the SVM classifier on the Aurora data set with a labeling threshold of $T_{60} = 0.7$ seconds (indicated by the dashed line). The overall accuracy on this data is 80%. The lines indicate the percentage of sound samples that were classified as either below the threshold ($T_{60} < 0.7$) or above the threshold ($T_{60} > 0.7$) for each reverberation level.

Conditions outside of ASR. Methods to assess or resolve the effects of reverberation should be tested in real-world environments, that is, with continuous audio recordings in unknown and possibly variable conditions. When the experimental paradigm is shifted toward these challenging conditions, the development of techniques for audio quality improvement needs to focus on robustness instead of perfection in limited domains.

2.5 Conclusion

In this chapter we have explained that automatic sound recognition in real-world environments requires sound event segregation that works in variable environ-
ments, that is, environments with different sorts and levels of transmission effects and varying co-occurring sounds. Furthermore, we have discussed that current techniques for automatic sound recognition are mostly designed for specific applications, such as speech recognition, and not for general purpose sound event recognition in unconstrained environments. In other words, the semantics of the sound are given by the application. Therefore, they can rely on signal-driven methods (based on the acoustic properties of the signal) to deal with transmission effects within a single domain. However, in real-world environments the semantics or context of sound events are not stable or even known. Instead of assuming a specific context, we argue that the context can be learned, and used to manage unreliable signal information. In fact, robust general purpose sound event recognition is infeasible with signal processing techniques only, because they can at most provide hypotheses of components that are likely to constitute a single sound event. Selecting and identifying the target sound event is only possible by means of non-acoustic factors, such as the goal of a system.

Moreover, even if researchers can assume or have accomplished perfect sound event segregation and classification for real-world sounds in a similar way as in speech and music processing, the meaning of the classified event is not yet known. When people listen in everyday life, they give meaning to the events that they hear. This meaning is not only based on acoustic properties or class membership. People use their memory, experience, and expectancy to give meaning to their (sonic) environment. Hence, these factors influence what is heard (segregated). In the next chapter we discuss the formation of the human percept of a sound event, and the role of context in this process.