ACOUSTIC EVENT DETECTION USING SIGNAL ENHANCEMENT AND SPECTRO-TEMPORAL FEATURE EXTRACTION

Jens Schröder[1], Benjamin Cauchi[1], Marc René Schädler[2], Niko Moritz[1], Kamil Adiloglu[3], Jörn Anemüller[1,2], Simon Doclo[1,2], Birger Kollmeier[1,2,3], Stefan Goetze[1]

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]University of Oldenburg, Department of Medical Physics and Acoustis, Oldenburg, Germany
[3]Hörtech gGmbH, Oldenburg, Germany

ABSTRACT

In this paper, an acoustic event detection system is proposed. It consists of a noise reduction signal enhancement step based on the noise power spectral density estimator proposed in [1] and on the noise suppression by [2], a Gabor filterbank feature extraction stage and a two layer hidden Markov model as back-end classifier. Optimization on the development set yields up to a F-Score of 0.73 on frame based and 0.63 on onset and offset based measure.

Index Terms— acoustic event detection, Gabor filterbank, minimum statistics

1. INTRODUCTION

Acoustic event detection (AED) is increasingly used in many fields of application, e.g. for surveillance and security issues [3–6] or in the field of Ambient Assisted Living (AAL) [7, 8]. In a past AED challenge, i.e. the CLEAR'07 (Classification of Events, Activities and Relationships) challenge [9] that was part of the CHIL project [10], detecting acoustic events in a meeting room scenario has been addressed. The proposed AED approaches were mainly based on Mel frequency cepstral coefficients (MFCCs) in conjunction with hidden Markov models (HMMs) [11–13]. Only one approach utilized a support vector machine instead [14]. The AED system that could demonstrate best recognition performance in the CLEAR'07 challenge used different feature streams in conjunction with a feature selection algorithm and a HMM back-end [15].

In this contribution, a system is proposed that can be separated into three main processing blocks (cf. Figure 1(a)). First, the acoustic input signal is preprocessed using minimum statistics (MS) noise estimation [1] and log-amplitude spectral attenuation for noise suppression [2]. Second, acoustic features are extraced by applying 2D Gabor filters on a Mel-warped spectro-temporal representation [16]. Third, the feature stream is fed to an HMM back-end. The performance of the proposed AED system is evaluated using the office live environment recordings of the IEEE AASP challenge [17].

This work was partially funded by the DFG Cluster of Excellence 1077 "Hearing4all" and by the Federal Ministry of Education and Research, BMBF, (project AALADIN,'V4PFL013).

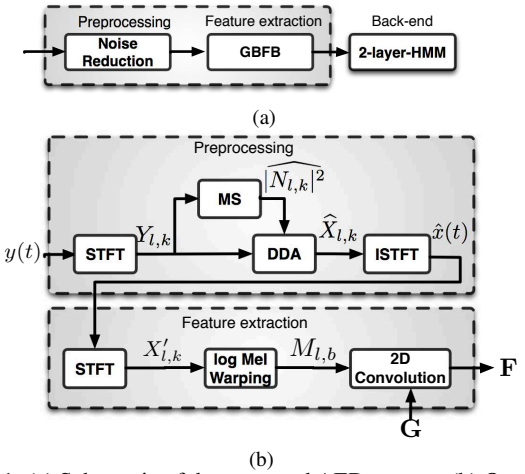

Figure 1: (a) Schematic of the proposed AED system. (b) Overview of preprocessing and feature extraction.

2. PREPROCESSING AND FEATURE EXTRACTION

The time-domain input signal $y(t) = x(t) + n(t)$ consists of the signal $x(t)$ containing the events of interest and additive noise $n(t)$. In this contribution, the acoustic input signal y(t) is resampled to $f_s = 16$ kHz and one channel is used. By short time Fourier transform (STFT) using a Hanning window of 32 ms and 50% overlap we obtain $Y_{l,k} = X_{l,k} + N_{l,k}$ in the block frequency domain. Here, l and k are the frame and frequency bin indices of the complex spectrum, respectively. The noise-reduction consists of two steps as shown in Figure 1(b), a noise power spectral density (PSD) estimator and a noise suppression. The MS estimator as decribed in [1] estimates the noise PSD from a noisy signal by tracking the minima of the power of the input signal for each frequency bin along time. As the office noise of this challenge's corpus is only slowly time-varying, MS is an appropriate method to estimate the noise PSD $\widehat{|N_{l,k}|^2}$. The decision directed approach (DDA) [2] is used to remove the noise from the input spectrum $Y_{l,k}$ to obtain an estimate of the clean event signal $\widehat{X}_{l,k}$, and after inverse short time Fourier transform (ISTFT) the time-domain signal of the acoustic events $\hat{x}(t)$. As the feature extraction is also done in the frequency domain this ISTFT step is only made necessary by our implementation that uses different window length and overlap (25 ms and 15 ms) for the

feature extraction.

The processed signal $\hat{x}(t)$ is represented by spectro-temporal modulation patterns called Gabor filterbank (GBFB)-features as proposed in [16]. The use of Gabor filters is motivated by their similarity to spectro-temporal patterns of neurons in the auditory cortex of mammals [18] and it has been shown that GBFB-features can improve the robustness of classification schemes, e.g. for automatic speech recognition systems [19]. The features are extracted using the reference MATLAB implementation available online [20], STFT is applied to $\hat{x}(t)$ to obtain $X'_{l,k}$ which is warped to obtain the log Mel-spectrogram $M_{l,b}$, with b the Mel band index. In contrast to the GBFB reference implementation that uses 23 Mel-bands between 64 Hz and 4 kHz, here the frequency range is extended to 8 kHz and the number of Mel-bands set to 31. The resulting log Mel-spectrogram $M_{l,b}$ is then 2D-convolved with 2D-filterbanks $\mathbf{G}$ that are sensitive to frequency changes over time. The selected filterbanks $\mathbf{G}$ are shown on Figure 2. While purely spectral filters ($\omega_n = 0$) are sensitive to spectral patterns like tonal components, purely temporal filters ($\omega_k = 0$) are sensitive to broad-band onsets, where ω_n and ω_k are the modulation frequency over time ad over frequency respectively. The resulting feature matrix $\mathbf{F}$ of each data file is then fed to the HMM recognizer.

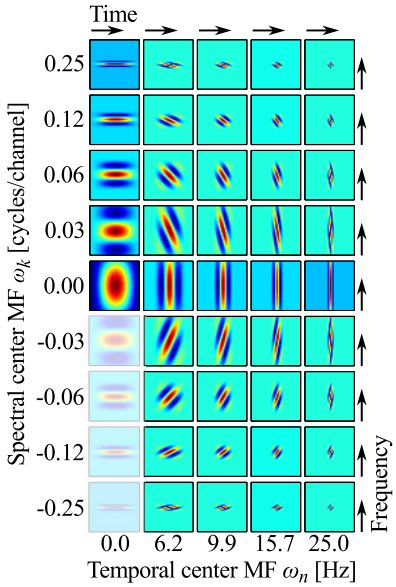

Figure 2: Shapes of the different 2D GBFB filters.

3. BACK-END CLASSIFIER

For the back-end classifier, the Hidden Markov Toolkit (HTK) [21] is applied to build up an HMM recognition network with a task grammar. HTK provides a speech recognition network of three levels: word level, model level and HMM level. In this contribution, events are treated like words. The model level, that is used in speech recognition to represent sub-words like phonemes, is not employed here. Thus, the whole recognizer can be interpreted as a two-layer HMM. The first layer is a fully connected HMM in which each state is an event, i.e. each event can be accessed at every time. The observations of these event states are themselves HMMs that are trained independently using the extracted features. These events are modeled by left-to-right HMMs with 3 emitting states (cf. [22]). To estimate time regions in a signal in which no active event is present, an extra *silence* class is modeled. For this class, 1 emitting state is implemented resulting in a simple Gaussian mixture model (GMM). The number of Gaussian mixtures for the event classes $\mathcal{M}_{\text{ev}}$ and for *silence* $\mathcal{M}_{\text{sil}}$ are adjusted on the development set.

To estimate the time regions of events in a signal, Viterbi decoding [21] is used. Since the output can be highly fragmented, i.e. several insertion and deletion errors may occur, a fixed logarithmic probability insertion penalty p is added to every event state transition [21]. Thus, the probability to remain in an event/*silence* state can be increased and a less scattered output is achieved.

4. EXPERIMENT SETUP

A training set and a development set of office recordings called Office Live Recordings (OL) were published by the organizers of the AED challenge [17]. The final testing set was kept secret and will be used for evaluation by the organizers. The published database consists of stereo recordings made in an office environment at 44.1 kHz sampling frequency. Although recordings from a 4 channel audio recording device are available they are not used for this contribution. The recordings comprise 16 classes: *door knock, door slam, speech, human laughter, clearing throat, coughing, drawer, printer, keyboard clicking, mouse click, pen dropping, switch, keys, phone ringing, alert, page turning*. The given training set contains 20 to 24 single trimmed recordings per class with small silent margins in the beginning and ending. The development set covers three recordings with altogether 110 events in continuous streams, i.e. single events alternated with short pauses.

As evaluation measures the F-Score and the acoustic event error rate (AEER) are used on frame, event onset, event on-/offset and class-wise level [17]. The F-Score F represents the relation between the precision P and the recall R.

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

The AEER is the sum of insertions I, deletions D and substitutions S relative to the number of reference events N.

$$\text{AEER} = \frac{I + D + S}{N} \tag{2}$$

5. RESULTS

The mentioned parameters for number of Gaussian mixture components $\mathcal{M}_{\text{ev}}$ and $\mathcal{M}_{\text{sil}}$ and the insertion penalty p for the back-end HMM are optimized on the development set. The number of mixtures is kept equal for all events except for *silence* where a different number is possible.

Numbers between 1 and 8 have been tested. This is done once for the frame as well as for the on-/offset based measure. The best results, optimized for the F-Score are shown in Table 1. If optimized for the frame based performance, the F-Score is of 0.76 when the on-/offset one is of 0,63. On the other hand if optimized for the on-/offset performance, the F-Score is of 0.73 when the on-/offset one is of 0,66.

6. CONCLUSION

In this contribution, we propose an AED system that applies a noise reduction based on a MS noise estimation and a DDA subtraction before representing the signal with GBFB features. Those features

Table 1: F-Score and AEER for the frame, onset and on-/offset based mesures (columns). The lines indicate whether the settings were optimized due to the frame based F-Score or the on-/offset one.

	settings			frame		onset		on-/offset	
optimization	$\mathcal{M}_{\text{ev}}$	$\mathcal{M}_{\text{sil}}$	p	F	AEER	F	AEER	F	AEER
frame	3	8	-700	0.76	0.69	0.63	0.92	0.60	1.00
on-/offset	2	8	-700	0.73	0.67	0.66	0.75	0.63	0.80

are then fed to a two-layer-HMM that detects and classifies events. Parameter optimization for the number of Gaussian mixtures and a penalty p are done on the development set. The optimization is done on the F-Score of the frame based approach as well as on the on-/offset one which leads to slightly different results.

7. APPENDIX

The AED system is implemented in MATLAB utilizing the *Signal Processing Toolbox*. The code is run by

`run_classifier(`*input*`,` *output*`,` *bFramemode*`)`

input	...	path of input wav file.
output	...	path of output recognition file.
bFramemode	...	boolean indicating if frame based (1) metric or any event based metric (0) is to be evaluated.

The algorithm includes HTK [21], that is compiled for a Linux 64-bit system. If another operating system is used the HVite function has to be replaced.

8. REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] D. P. W. Ellis, "Detecting alarm sounds," in *Proceedings of the Recognition of real-world sounds: Workshop on consistent and reliable acoustic cues*, Aalborg, Denmark, 2001, pp. 59–62.

[4] P. W. van Hengel, M. Huisman, and J.-E. Appell, "Sounds like trouble," in *Human Factors - Security and Safety*, D. de Waard, J. Godthelp, F. Kooi, and K. Brookhuis, Eds. Shaker Publishing, Maastricht, The Netherlands., 2009, pp. 369–375.

[5] R. A. Lutfi and I. Heo, "Automated detection of alarm sounds," *Journal of the Acoustical Society of America*, vol. 132, no. 2, Sep. 2012.

[6] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic Acoustic Event Detection in Traffic Noise by Part-Based Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[7] J. Schröder, S. Wabnik, P. van Hengel, and S. Goetze, "Detection and Classification of Acoustic Events for In-Home Care," in *Ambient Assisted Living*. Springer, 2011, pp. 181 – 195.

[8] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice Activity Detection Driven Acoustic Event Classification for Monitoring in Smart Homes," in *3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Rome, Italy, Nov. 2010.

[9] CLEAR: Classification of Events, Activities and Relationships, http://clear-evaluation.org/?CLEAR, 2007.

[10] CHIL: Computers in the human interaction loop, http://chil.server.de/.

[11] C. Boukis and L. C. Polymenakos, "The Acoustic Event Detector of AIT," in *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds. Springer, 2008, vol. 4625, pp. 328–337.

[12] C. Zieger, "An HMM Based System for Acoustic Event Detection," in *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds. Springer, 2008, vol. 4625, pp. 338–344.

[13] T. Heittola and A. Klapuri, "TUT Acoustic Event Detection System 2007," in *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds. Springer, 2008, vol. 4625, pp. 364–370.

[14] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic Event Detection: SVM-Based System and E0valuation Setup in CLEAR'07, editor = Rainer Stiefelhagen and Rachel Bowers and Jonathan G. Fiscus, booktitle = Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers, publisher = Springer, series = Lecture Notes in Computer Science, volume = 4625, pages = 354-363, year = 2008, isbn = 978-3-540-68584-5,."

[15] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. S. Huang, "HMM-Based Acoustic Event Detection with AdaBoost Feature Selection," in *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds. Springer, 2008, vol. 4625, pp. 345–353.

[16] M. R. Schädler and B. Kollmeier, "Normalization of spectro-temporal gabor filter bank features for improved robust automatic speech recognition systems," in *Proceedings of Interspeech*, Portland, USA, 2012.

[17] IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events, http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/, 2013.

[18] A. Qiu, C. E. Schreiner, and M. A. Escabí, "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[19] N. Moritz, M. R. Schädler, K. Adiloglu, B. T. Meyer, T. Jürgens, T. Gerkmann, B. Kollmeier, S. Doclo, and S. Goetze, "Noise Robust Distant Automatic Speech Recognition Utilizing NMF Based Source Separation and Auditory Feature Extraction," in *CHiME challenge workshop 2013*, Vancouver, Canada, June 2013.

[20] M.R. Schädler: GBFB feature extraction reference implementation, http://medi.uni-oldenburg.de/GBFB, 2012.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 2006.

[22] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, Aug. 2010, pp. 1267–1271.