ON THE USE OF SPECTRO-TEMPORAL FEATURES FOR THE IEEE AASP CHALLENGE 'DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS'

Jens Schröder[1]*, Niko Moritz[1], Marc René Schädler[2], Benjamin Cauchi[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 Acoustics, Oldenburg, Germany
[3]Hörtech gGmbH, Oldenburg, Germany

ABSTRACT

In this contribution, an acoustic event detection system based on spectro-temporal features and a two-layer hidden Markov model as back-end is proposed within the framework of the IEEE AASP challenge 'Detection and Classification of Acoustic Scenes and Events' (D-CASE). Noise reduction based on the log-spectral amplitude estimator by [1] and noise power density estimation by [2] is used for signal enhancement. Performance based on three different kinds of features is compared, i.e. for amplitude modulation spectrogram, Gabor filterbank-features and conventional Mel-frequency cepstral coefficients (MFCCs), all of them known from automatic speech recognition (ASR).

The evaluation is based on the office live recordings provided within the D-CASE challenge. The influence of the signal enhancement is investigated and the increase in recognition rate by the proposed features in comparison to MFCC-features is shown. It is demonstrated that the proposed spectro-temporal features achieve a better recognition accuracy than MFCCs.

Index Terms— acoustic event detection, Gabor filterbank, amplitude modulation spectrogram, IEEE AASP D-CASE challenge

1. INTRODUCTION

Acoustic event detection (AED) is increasingly used in various application fields, e.g. for surveillance and security. Examples include detection and classification of emergency situations in public environments such as siren detection [3] and recognition of screams [4] as well as health monitoring, e.g. respiratory sound monitoring [5]. Another application for AED is the improvement of speech recognition systems, e.g. for meeting-room and seminar situations. In this context, the well known CLEAR'07 (classification of events, activities and relationships) challenge [6] that was initiated by the CHIL (computers in the human interaction loop) project [7] was organized, addressing detection of acoustic events in a meeting room scenario. The AED approaches proposed for this challenge were mainly based on Mel-frequency cepstral coefficient (MFCC)-features and hidden Markov model (HMM) classifiers [8]. Only one approach utilized a support vector machine (SVM) instead. 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. This approach was further improved leading to a tandem connectionist-HMM classifier, which combined the benefit of discriminative artificial neural networks (ANNs) with the segmentation capabilities of HMMs [9]. The output was re-scored by a SVM-GMM-supervector [10] approach.

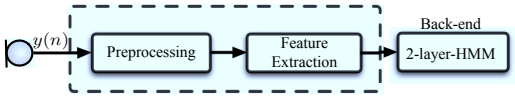

Figure 1: Schematic of the proposed AED system.

A recent approach described in [11] also utilized a tandem connectionist-HMM but focuses on the front-end. Spectro-temporal harmonic percussive sound separation (HPSS)-features known from music processing were adopted for AED. HPSS differentiates sounds by the temporal and spectral smoothness.

In [12], spectro-temporal features were adopted by applying non-negative matrix factorization (NMF). This approach led to a codebook of spectro-temporal patches that described events. The segmentation was done by an HMM. In this contribution, two psycho-physiological motivated spectro-temporal features, that have been recently used in automatic speech recognition (ASR), are investigated. The classifying system proposed can be separated into three main processing blocks (cf. Figure 1). Firstly, the acoustic input signal is preprocessed by a log-amplitude spectral attenuation for noise reduction (NR) [1] with a minimum statistics (MS) noise estimator [2]. Secondly, acoustic features are extracted. One approach utilizes Gabor filterbank (GBFB)-features that exploit spectro-temporal information by applying 2D Gabor filters on a Mel-warped time-frequency representation [13]. The second approach exploits amplitude modulations from a Mel-warped time-frequency representation to calculate the amplitude modulation spectrogram (AMS). Both features that are newly proposed for the task of AED in this contribution are compared to standard MFCCs with additional time derivatives of first (Δ) and second ($\Delta\Delta$) order. Finally, the feature stream is fed to an HMM back-end. The performance of the proposed AED system is evaluated with the office live recordings provided by the IEEE AASP challenge 'Detection and Classification of Acoustic Scenes and Events' (D-CASE). The influence of the preprocessing step is tracked.

2. PREPROCESSING

The time-domain input signal $y(n) = x(n) + d(n)$ consists of the signal $x(n)$ containing only the events of interest and an additive

*This work was partially funded by DFG Cluster of Excellence 1077 Hearing4all, DFG FOR-1732 and European Commision (Project EAR-IT (No. 318381)).

978-1-4799-0972-8/13/$31.00 ©2013IEEE

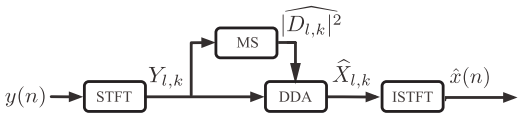

Figure 2: Overview of the preprocessing.

noise disturbance $d(n)$. In this contribution, the acoustic input signals $y(n)$ from [14] are resampled to $f_s = 16$ kHz and only one channel is selected. By short-time Fourier transform (STFT) using a Hann window of 32 ms and 50% overlap $Y_{\ell,k} = X_{\ell,k} + D_{\ell,k}$ in the block frequency domain is obtained. Here, ℓ and k are the frame and frequency bin indices of the complex spectra, respectively. The noise reduction (NR) consists of two steps, a noise power spectral density (PSD) estimator [2] and a spectral enhancement. As the office noise of this challenge's corpus is only slowly time-varying, a minimum statistics (MS) approach is used to estimate $\widehat{|D_{\ell,k}|^2}$. The decision-directed approach (DDA) [1] is used for NR to obtain the estimate $\widehat{X}_{\ell,k}$ Finally, the time domain signal $\hat{x}(n)$ is calculated using the inverse short-time Fourier transform (ISTFT). An overview over the NR procedure is depicted in Figure 2.

3. FEATURE EXTRACTION

To extract relevant information from a signal, it is transformed to a feature domain. In the following, three different kinds of features are tested that all work on the same spectro-temporal representation. Hence, the sampled signal $y(n)$ is framewise processed by a Hamming window of size $N = 400$ samples, i.e. each frame ℓ comprises 25 ms. The frames are shifted by n_s = 160 samples, i.e. 10 ms. The frames are transformed to frequency domain by a discrete Fourier transformation (DFT). The absolute value $|Y_{\ell,k}|$ of the resulting spectrogram is Mel-warped by triangular shaped Mel-filters $F_{k,m}$ in a frequency region between 64 Hz and 8 kHz and logaritmized leading to a log-scaled Mel-spectrogram with Mel-bands m

$$\tilde{Y}_{\ell,m} = \log\left(\sum_{k=0}^{N-1} |Y_{\ell,k}| \cdot F_{k,m}\right) \qquad 0 \le m \le M-1, \quad (1)$$

with $M = 31$ representing the number of Mel-filters.

For comparison, standard MFCCs are adopted. A discrete cosine transform (DCT) transforms the log Mel-spectrogram $\tilde{Y}_{l,m}$ to the cepstral domain, i.e.

$$\check{Y}_{\ell,c} = \sum_{m=0}^{M-1} \tilde{Y}_{\ell,m} \cos\left(\frac{\pi}{M}\left(m+\frac{1}{2}\right)c\right) \quad 0 \le c \le C-1, \quad (2)$$

with the number of cepstral coefficients $C \le M$. For MFCC evaluation, only the first $C = 12$ coefficients including the 0th DC-coefficient, that represents the signal energy, are further employed. The first derivatives Δ are calculated on a time scale of 65 ms, i.e. 5 frames, the second derivatives $\Delta\Delta$ are calculated on a time scale of 105 ms.

3.1. Amplitude modulation spectrogram (AMS)

The AMS is constructed using a filter set that extracts temporal amplitude modulation frequency components in subbands of a spectro-temporal representation [15]. In this study, it is computed based on $\tilde{Y}_{\ell,m}$. A DCT along the acoustic frequency axis is employed in addition to filtering of temporal trajectories by the AMS filter set. Thus, the AMS processing is effectively conducted on the cepstrogram $\check{Y}_{\ell,c}$ of the signal. The AMS filter set employed here consists of five constant-Q amplitude modulation (AM) filters. The filters are a product of a Hann window

$$h_b(\ell) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi\ell}{b}\right) & -\frac{b}{2} < \ell < -\frac{b}{2}, \\ 0 & \text{else}, \end{cases} \quad (3)$$

where b denotes the width of the envelope, mutiplied by an sinusoidal carrier function with frequency ω

$$s_\omega(\ell) = \exp(j\omega\ell). \quad (4)$$

They can be described by frame index ℓ; the central time frame ℓ_0, the temporal modulation frequency ω and the number of semi cycles under the envelope ν, hence

$$q(\ell_0, \omega, \ell, \nu) = s_\omega(\ell - \ell_0) \cdot h_{\frac{\nu}{2\omega}}(\ell - \ell_0). \quad (5)$$

The applied modulation frequencies ω that have been optimized for ASR are 0 Hz, 3.125 Hz, 6.25 Hz, 12.5 Hz and 25 Hz [15]. The AM filters have zero mean except for the DC filter $\omega = 0$. The AMS-features are calculated by convolution of the cepstrogram $\check{Y}_{\ell,c}$ with the real parts of the filters, i.e.

$$Q_{\ell,c}(\ell_0, \omega, \nu) = \sum_\lambda \check{Y}_{\lambda,c} \Re\{q(\ell_0, \omega, \lambda + \ell, \nu)\}. \quad (6)$$

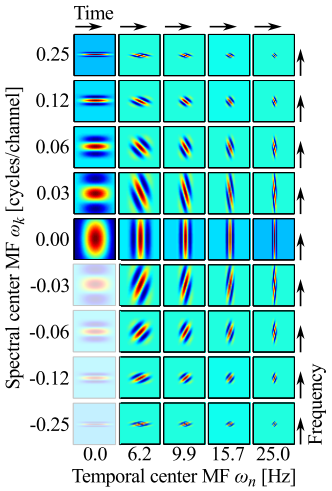

Figure 3: Shapes of the different 2D GBFB filters.

3.2. Gabor filterbank (GBFB)-features

The signal $y(n)$ can also be represented by spectro-temporal modulation patterns called GBFB-features as proposed in [13]. The use of Gabor filters is motivated by their similarity to spectro-temporal patterns of neurons in the auditory cortex of mammals [16] and it has been shown that GBFB-features can improve the robustness of automatic speech recognition systems [17]. A Gabor filter is a product of a 2D Hann-shaped envelope function with an 2D sinusoidal carrier. Thus, a Gabor filter can be expressed by the frequency and frame indicies m and ℓ, the central frequency channel m_0 and the

central time frame ℓ_0, the spectral and temporal modulation frequencies ω_m and ω_ℓ and the number of semi cycles under the envelope ν_m and ν_ℓ, i.e.

$$\begin{aligned} g\left(m_0, \ell_0, \omega_m, \omega_\ell, m, \ell, \nu_m, \nu_\ell\right) \\ = s_{\omega_m}\left(m - m_0\right) \cdot s_{\omega_\ell}\left(\ell - \ell_0\right) \\ \cdot h_{\frac{\nu_m}{2\omega_m}}\left(m - m_0\right) \cdot h_{\frac{\nu_\ell}{2\omega_\ell}}\left(\ell - \ell_0\right). \end{aligned} \quad (7)$$

In Figure 3, the shapes of the filterbank are plotted. To construct the features, the log-scaled Mel-spectrogram $Y_{\ell,m}$ is filtered with the real parts of the filters that are sensitive to frequency changes over time,

$$\begin{aligned} & G_{\ell,m}\left(m_0, \ell_0, \omega_m, \omega_\ell, \nu_m, \nu_\ell\right) \\ & = \sum_{\mu} \sum_{\lambda} \tilde{Y}_{\lambda,\mu} \Re\left\{g\left(m_0, \ell_0, \omega_m, \omega_\ell, \mu + m, \lambda + \ell, \nu_m, \nu_\ell\right)\right\}. \end{aligned} \quad (8)$$

The maximum filter size is limited to 69 filter channels and 40 time frames corresponding to 415 ms. The used filterbanks are shown in Figure 3. While purely spectral filters ($\omega_\ell = 0$) are sensitive to spectral patterns like tonal components, purely temporal filters ($\omega_m = 0$) are sensitive to broad-band onsets.

4. BACK-END

For the back-end, the Hidden Markov Toolkit (HTK) [18] 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 where 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 on the extracted features. These events are modeled by left-to-right HMMs with 3 emitting states. To estimate time regions in a signal where 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. Diagonal covariance matrices are applied since the training set is small (curse of dimensionality).

To estimate the time regions of events in a signal, Viterbi decoding [18] 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 can be added to every event state transition [18]. Thus, the probability to remain in an event/*silence* state can be increased and a less scattered output is achieved.

5. EXPERIMENTS AND EVALUATION

A training and a development set with office live recordings (OL) were published within the D-CASE challenge [14]. The final testing set was kept secret and evaluated by the organizers. The published datasets consist 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, only one channel is 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 at 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.

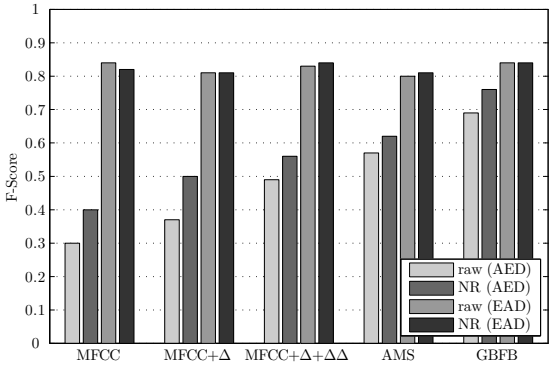

Figure 4: F-Score of the the general classification AED and the EAD with noise reduction (NR) and without (raw).

As evaluation measures the F-Score and the acoustic event error rate (AEER) are used. The F-Score F represents the relation between the precision P and the recall R

$$P = \frac{N_{\text{corr}}}{N_{\text{est}}}; \quad R = \frac{N_{\text{corr}}}{N_{\text{ref}}}; \quad F = \frac{2 \cdot P \cdot R}{P + R}, \quad (9)$$

where N_{corr} denotes the number of correct hits, N_{est} the number of estimated events and N_{ref} the number of reference events. The AEER is the sum of insertions I, deletions D and substitutions S relative to the number of reference events N, i.e.

$$\text{AEER} = \frac{I + D + S}{N}. \quad (10)$$

These measures are used on frame level for frames of 10 ms duration. The parameters for the proposed algorithm, i.e. the number of Gaussian mixture components $\mathcal{M}_{\text{ev}}$ and $\mathcal{M}_{\text{sil}}$ and the insertion penalty p for the back-end HMM have been 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 were tested. The relevant optimization score is the F-Score F.

Two tasks have been evaluated. The first task is the general AED task, i.e. to segment and classify each event individually by its unique HMM. The second task, called EAD according to the more common term voice activity detection (VAD) in speech recognition, is to segment any event in a stream. Therefore, all events, except for *silence* that is treated separately, are trained into one HMM.

The best F-Score and AEER on the development set for both tasks and for different features are shown in Figures 4 and 5, respectively. Not surprisingly, the EAD-task results in higher F-Score and lower AEERs than the more complex AED task since the first one is only detecting events whereas the latter is a combination of segmentation and classification. All tested features perform more or less equal in detecting events. The applied preprocessing only has

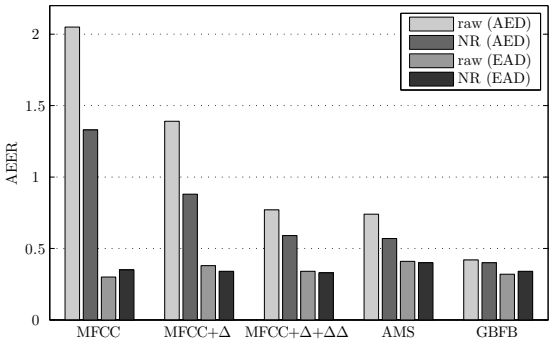

Figure 5: AEER of the the general classification AED and the EAD with noise reduction (NR) and without (raw).

minor influence on this task. However, for AED, the preprocessing leads to higher F-Score and lower AEER for every tested feature in comparison to the unpreprocessed data.

MFCCs without any derivatives, hence just working on the short-time spectrum, yield a poor performance. The more temporal context is exploited by adopting derivatives of first (65 ms) and second order (105 ms), the more robust the recognition becomes. GBFB and AMS that operate on a time span of up to $\approx$ 415 ms achieve the highest performance.

Though AMS-features work on a similar time scale as GBFB-features they result in a slightly lower recognition performance. A reason for this might be that AMS-features only operate on the time and frequency domain independently, i.e. they are comparable to Gabor filters with either $\omega_\ell = 0$ or $\omega_m = 0$. GBFB-features on the other hand exploit the spectro-temporal information jointly.

The classifier using NR and GBFB-features was evaluated by the challenge organizers applying the hidden testing set. The performance decreased from 76% to 62% in F-Score. This might be due to slight overfitting. However, since for creating the event models the training set, that is conjunct to the development set, was used, it is not likely that the features themselves cause overfitting, though they comprise 80 dimensions for AMS and 455 for GBFB what is often experienced to lead to such effects. Only the number of Gaussian mixtures $\mathcal{M}_{\text{ev}}$ and $\mathcal{M}_{\text{sil}}$ and the insertion penalty p were adjusted on the development set and could hence lead to overfitting effects. Testing on the development set, the F-Score is indeed decreased from 76% to 67% if $\mathcal{M}_{\text{ev}}$ is increased from 3 to 4 which could be a hint for overfitting. But this also applies even more for MFCCs (incl. Δ and $\Delta\Delta$) for which the F-Score decreases from 54% to 27% by increasing $\mathcal{M}_{\text{ev}}$ by one. Since the testing set is not available yet, it could not be evaluated if the other features suffered similar degradation in performance like GBFB-features on that set.

6. CONCLUSION

In this contribution, an AED system is proposed that firstly applies signal enhancement to an acoustic signal. Secondly, MFCC and spectro-temporal features known from ASR, i.e. GBFB and AMS, are extracted and fed to a two-layer-HMM. It was shown that all features perform similarly for the event segmentation task and the proposed preprocessing is not beneficial there. However, for segmentation and classification, noise reduction results in better performance for all features. Both newly adopted features for the task of AED performed better than standard MFCCs, where GBFB performed best. In future work, the parameterization for the features, which is adopted from ASR, may need to be adjusted more properly to the new application area.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[3] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic Acoustic Event Detection in Traffic Noise by Part-Based Models," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 493 – 497.

[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] F. Jin, F. Sattar, and S. Krishnan, "Log-frequency spectrogram for respiratory sound monitoring," in *Proc. ICASSP*, Mar. 2012, pp. 597–600.

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

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

[8] R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds., *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 4625.

[9] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[10] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 97–100.

[11] M. Espi, M. Fujimoto, D. Saito, N. Ono, and S. Sagayama, "A tandem connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4293–4296.

[12] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. WASPAA*, Mohonk, USA, October 2011, pp. 69–72.

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

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

[15] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth," in *Proc. Interspeech*, Portland, USA, Sep. 2012.

[16] 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.

[17] 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 *2nd CHiME challenge workshop 2013*, Vancouver, Canada, Jun. 2013.

[18] 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.