Acoustic Scene Classification using Time-Delay Neural Networks and Amplitude Modulation Filter Bank Features

Niko Moritz, Jens Schröder, Stefan Goetze [1]

Fraunhofer IDMT, Project Group for Hearing,
Speech, and Audio Technology
Marie-Curie-Str. 2, 26121 Oldenburg, Germany
niko.moritz@idmt.fraunhofer.de

Jörn Anemüller, Birger Kollmeier

University of Oldenburg
Medizinische Physik & Hearing4all
Carl-von-Ossietzky-Str. 9-11
26129 Oldenburg, Germany

ABSTRACT

This paper presents a system for acoustic scene classification (SC) that is applied to data of the SC task of the DCASE'16 challenge (Task 1). The proposed method is based on extracting acoustic features that employ a relatively long temporal context, i.e., amplitude modulation filer bank (AMFB) features, prior to detection of acoustic scenes using a neural network (NN) based classification approach. Recurrent neural networks (RNN) are well suited to model long-term acoustic dependencies that are known to encode important information for SC tasks. However, RNNs require a relatively large amount of training data in comparison to feed-forward deep neural networks (DNNs). Hence, the time-delay neural network (TDNN) approach is used in the present work that enables analysis of long contextual information similar to RNNs but with training efforts comparable to conventional DNNs. The proposed SC system attains a recognition accuracy of 76.5 %, which is 4.0 % higher compared to the DCASE'16 baseline system.

***Index Terms*—** Time-delay neural networks, acoustic scene classification, DCASE, amplitude modulation filter bank features.

1. INTRODUCTION

Machine listening for automatic scene classification (SC) becomes increasingly popular, e.g., as reflected by a past SC challenge that compared research results of many international research teams [1]. Devices like hearing-aids, smart-phones, and robotic platforms are equipped with microphones and applications analyzing the acoustical environment, e.g., to allow for switching parameters of signal processing schemes [2,3]. Hence, in many situations it is of interest to know the environment in which an electronic device is used, e.g., to distinguish acoustic conditions of a conference room, cafeteria or subway. Acoustically driven scene classification (SC) algorithms aim at classifying the surrounding environment automatically by identifying acoustic events and sound characteristics that are specific for the environment. In contrast to acoustic event detection (AED) [4,5,6], individual events are of minor interest and since acoustic scenes do not change rapidly, constraints on temporal resolution for SC are more relaxed than for AED and often comprise lengths of 30 seconds [1,7,8] up to 3 minutes [9].

Different approaches have been proposed for the purpose of automatic SC such as the use of a bag-of frames approach [9], for which a Gaussian mixture model (GMM) in combination with Mel-frequency cepstral coefficients (MFCCs) are adopted. This approach has established itself in the field of scene classification and till today is still accepted as a reasonable baseline system for the DCASE challenges 2013 [1] and 2016 [7], though most of the systems in the DCASE'13 challenge could outperform the baseline results. Proposed features within DCASE'13 ranged from standard features such as MFCCs [10,11] and low-level features like energy, spectral flux etc. [12,13] over cochleograms [14] to histogram of gradients (HOG) features [8] and Gabor filter bank (GFB) features [15] that both have been derived from computer vision. Most back-end classifiers used for the DCASE'13 challenge were based on support vector machines (SVM) [16,12,14,8,11].

In a recent publication [17], the idea of using HOG features was revisited and improved by using them in conjunction with the subband power distribution (SPD). Other common approaches for SC apply non-negative matrix factorization (NMF) to spectrograms to decompose features before classification [18,19].

In this contribution, we propose the use of amplitude modulation filter bank (AMFB) features [20] in combination with a neural network (NN) based classifier for the task of SC. AMFB features analyze temporal amplitude fluctuations of static MFCCs within modulation frequency subbands. In combination with GMM and deep neural network (DNN) based systems, AMFB features have demonstrated to outperform numerous other common feature extraction methods in automatic speech recognition (ASR) [21,20,22]. In addition to AMFB features, spectral flux, spectral centroid, and spectral entropy features are calculated and appended.

DNNs are well established in, e.g., ASR [23,24] and have recently received increased attention also in the field of AED [25,26]. In ASR and AED, DNNs have proven to outperform conventional GMM-HMM approaches [27,25] and NMF-based features [26] under the constrained of availability of sufficient training data. Hence, DNNs may also be well suited for acoustic SC, since SC corpora mostly comprise several hours of data, e.g., the LITIS Rouen dataset [8] that comprises 25 hours of urban sound scenes, which is necessary to train a reasonable NN-based system.

Here, we report on our work on the DCASE'16 challenge and results are shown using a time-delay neural network (TDNN)

[1] This work was funded in parts by the European Commission (project EcoShopping, project no.609180) and the Federal Ministry for Education and Research (BMBF), project ACME 4.0, FKZ 16ES0469).

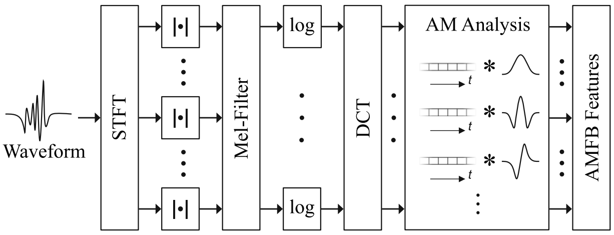

Fig. 1. Signal processing scheme to extract amplitude modulation filter bank features.

architecture [28] that relies on AMFB features as an input for the Task 1 of the DCASE'16 challenge, which comprises less than 10 hours of recordings [7]. Results are compared to the DCASE'16 baseline system that applies GMM acoustic models in combination with MFCC features.

2. METHODS

2.1. Extraction of Amplitude Modulation Filter Bank Features

The acoustic feature extraction scheme employs the amplitude modulation filter bank (AMFB) to decompose short-term spectral features into AM frequency components [20]. Signal processing steps are depicted in Fig. 1. The short-term spectral representation $Y_k(l)$ for block l is calculated by applying a discrete Fourier transform (DFT) on audio segments of 25 ms length with a hop size of 10 ms. Segments are windowed by the Hann function $w_b(n)$ to minimize the spectral leakage effect.

$$Y_k(l) = \sum_{n=-\infty}^{\infty} y(n) \cdot w_b(n-l) \cdot e^{-\frac{j2\pi kn}{N}} \quad ,0 \le k \le N-1 \qquad (1)$$

$$w_b(n) = \begin{cases} 0.5 - 0.5 \cdot \cos\left(\frac{2\pi n}{b}\right) & ,0 \le n \le b \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

In (1) and (2), n, k, b, and N represent the discrete time and frequency indices, the analysis window length, and the DFT length, respectively.
The magnitude of the complex valued spectrum $Y_k(l)$ is passed to the triangular-shaped Mel filters $F_{k,m}$ that integrate DFT bins into $M = 40$ critical spectral bands. Mel-spectral energies are compressed using a logarithmic function, whereby the log-Mel-spectrogram $\hat{Y}_m(l)$ is derived for each Mel band m.

$$\hat{Y}_m(l) = \log\left(\sum_{k=0}^{N-1} \left|Y_k(l)\right| \cdot F_{k,m}\right) \quad ,0 \le m \le M-1 \qquad (3)$$

Log-Mel-spectral energies are analyzed by a discrete cosine transform (DCT), which leads to the cepstrogram $\tilde{Y}_c(l)$ with C being the DCT length.

$$\tilde{Y}_c(l) = \sum_{m=0}^{M-1} \hat{Y}_m(l) \cdot \cos\left(\tfrac{\pi}{M}\left(m + \tfrac{1}{2}\right)c\right) \quad ,0 \le c \le C-1 \qquad (4)$$

Table 1. Center frequency (CF) and bandwidth (BW) parameters of the amplitude modulation filter bank.

i	0	1	2	3	4
CF [Hz]	0	5.5	10.15	15.91	27.03
BW [Hz]	8.25	5.5	6.13	8.27	19.52

Temporal dynamics of the cepstrogram are analyzed using the AMFB. The AMFB consists of I complex exponential functions $q_i(l_0)$, that are windowed by the zero-phase Hann envelope $W_i(l_0)$.

$$q_i(l_0) = e^{-j\Omega_i l_0 \cdot T} \cdot W_i(l_0) \quad ,0 \le i \le I-1 \qquad (5)$$

$$W_i(l_0) = \begin{cases} 0.5 + 0.5\cos\left(\frac{2\pi l_0}{B_i}\right) & ,-\left\lceil\frac{B_i-1}{2}\right\rceil < l_0 < \left\lceil\frac{B_i-1}{2}\right\rceil \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

$$B_i = \frac{9.06}{2\pi \cdot \beta_i \cdot T} \qquad (7)$$

B_i determines the AM filter length with the sampling period T. Ω_i and β_i are the angular AM frequency and the -3 dB AM filter bandwidth, respectively. Convolution of $q_i(l_0)$ and $\tilde{Y}_c(l_0)$ yields the AM frequency decomposition of the cepstrum.

$$Q_{c,i}(l) = \left(\tilde{Y}_c * q_i\right)(l) \qquad (8)$$

Center frequency (CF) and bandwidth (BW) settings of the employed AM filters are presented in Table 1, which are derived by an ASR study on finding optimal AMFB parameters using different ASR corpora [22]. The last step of AMFB feature extraction is the concatenation of real and imaginary AM filter outputs to form a feature vector. Note that the imaginary part of the DC filter is zero, and thus is not taken into account.

2.2. Other Features

Spectral *flux*, spectral *centroid*, and spectral *entropy* features are derived according to Eq. 9-11 and appended to AMFB features.

$$Centroid(l) = \frac{\sum_{k=0}^{N-1}(k+1) \cdot \left|Y_k(l)\right|}{\sum_{k=0}^{N-1}\left|Y_k(l)\right|} \qquad (9)$$

$$Flux(l) = \sum_{k=0}^{N-1}\left(\left|Y_k(l)\right| - \left|Y_k(l-1)\right|\right)^2 \qquad (10)$$

$$Entropy(l) = -\sum_{k=0}^{N-1}\left|Y_k(l)\right|^2 \cdot \log_2\left(\left|Y_k(l)\right|^2\right) \qquad (11)$$

These three feature types are used to measure the spectral "center of mass", the spectral "rate of change", and the spectral "complexity" [12,13].

Table 2. Multi-splicing configuration of the TDNN system. Numbers in brackets indicate frame indices that are spliced together at each neural net layer.

NN-Layer	Input context [frames]
1	[-6 ,0 ,4]
2	[-12 ,0 ,12]
3	[-24 ,0 ,24]
4	[-50 ,0 ,50]
5	[0]

2.3. Classification

Extracted features are fed into a time-delay neural network (TDNN) to extract further acoustic cues and to perform the classification task. The TDNN differs from a conventional DNN by the multi-splicing concept that enables an efficient way of modelling a large temporal context [28,29]. Multi-splicing denotes a method by which feature frames and intermediate DNN-layer outputs are time-delayed and stacked to form the input to an upstream neural network (NN) layer. Splicing configurations per NN-layer are presented in Table 2. For example, the splicing notation [-6, 0, 4] in the first NN-layer denotes that the current frame minus six, the current frame itself, and the current frame plus 4 are spliced together by stacking input feature frames. We do not splice consecutive frames in the first layer, since AMFB features are used as input that already capture a temporal context of +/- 13 time frames and, thus, consecutive AMFB feature frames have highly overlapping filter functions and a high redundancy, respectively. The same principle applies to outputs of deeper NN-layers that capture an increasing temporal context due to the previous splicing stages. In total the TDNN captures feature frames ranging from -92 to +90, which corresponds with the feature frame rate of 100 Hz to a total temporal context of approx. 1.8 seconds.

The TDNN training is based on the greedy layer-wise supervised training [30] and the layer-wise backpropagation algorithm [27], respectively. As nonlinear activation units we are using the p-norm function that effect a dimension reduction of NN-layer outputs that each consist of 576 neurons in our setup. For example, for a group of G neurons x_i the p-norm output y is being computed by Eq. 12 with $p = 2$ and G = 6.

$$y = \|x\|_p = \left(\sum_{i=1}^{G} |x_i|^p \right)^{1/p} \tag{12}$$

Thus, the output of each NN-layer is reduced from 576 to 96. The final TDNN output layer has 15 neurons representing the 15 acoustic scenes that need to be discriminated.

3. EXPERIEMNTAL SETUP

For evaluating the algorithms, the database provided within the DCASE'16 challenge is used [7]. It consists of 15 scene classes: *lakeside beach*, *bus*, *cafe/restaurant*, *car*, *city center*, *forest path*, *grocery store*, *home*, *library*, *metro station*, *office*, *urban park*, *residential area*, *train*, and *tram*. Each scene is composed of 39 minutes of stereo recordings at 44.1 kHz sampling frequency that are trimmed to 30 second files. The data is divided into four disjoint sets to conduct a four-fold cross-validation, where all files belonging to one specific time/location are part of one set. Evaluation is conducted file-wise applying the accuracy measure, i.e., the number of correctly classified files in ratio to the total number of files.

Table 3. Acoustic scene classification results of the DCASE'16 baseline system and the proposed TDNN-based system.

	Hitrate [%]	
Environment	Baseline	Proposed Method
Beach	69.3	79.5
Bus	79.6	56.4
Café/Restaurant	83.2	44.9
Car	87.2	96.2
City Center	85.5	88.5
Forest Path	81.0	98.7
Grocery Store	65.0	87.2
Home	82.1	76.9
Library	50.4	69.2
Metro Station	94.7	79.5
Office	98.6	76.9
Park	13.9	56.4
Residential Area	77.7	88.5
Train	33.6	64.1
Tram	85.4	84.6
Average	72.5	76.5

4. RESULTS

In order to artificially augment the number of training frames the left and right channel of the stereo audio data is used in addition to the mean of both channels. In the testing phase the TDNN output for each of these three audio tracks is computed and the detected acoustic scene within an audio test file is based on a majority vote across frames and audio tracks. Note that prior to feature extraction we resampled data of the DCASE'16 challenge to 16 kHz.

Results of the proposed method and the DCASE'16 baseline system are presented in Table 3. On average the TDNN system achieves an improvement of 4 % compared to the baseline. Particular strength can be noted for the environments *beach, car*, *forest path*, *grocery store*, *library*, *park*, *residential area*, and *train*. A decreased performance is found for the environments *bus*, *café/restaurant*, *home*, *metro station*, and *office*. Fig. 2 depicts the confusion matrix of the proposed classification system. It shows that some environments with relatively low recognition rates, i.e., *café/restaurant, bus*, *library*, *park*, and *train*, are mostly confused with similar or related environments such as *café/restaurant* > *grocery store*, *bus* > *tram/train*, *library* > *home*, *park* > *residential area*, and *train* > *tram/bus*.

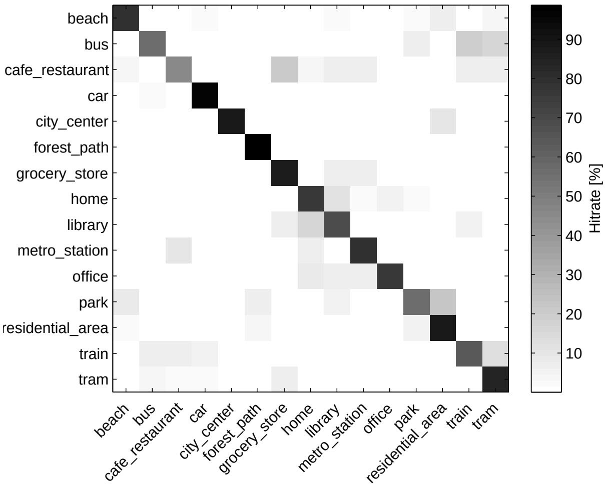

Fig. 2. Aggregate confusion matrix of the four-fold cross-validation sets. Rows are ground truths and columns recognized scenes.

5. DISCUSSION AND CONCLUSIONS

A time-delay neural network (TDNN) with amplitude modulation filter bank (AMFB) features plus spectral flux, centroid, and entropy features based acoustic scene classification approach is proposed that aims at analyzing a relatively long temporal context to identify acoustic environments. It is shown that the AMFB-TDNN system improves over a MFCC-GMM baseline system by approximately 4.0 %. Further improvements may be attained by additionally utilizing binaural cues of the stereo DCASE'16 data that is recorded using a manikin head with in-ear microphones and by emphasizing an even larger temporal context (the current system processes approx. 1.8 seconds of audio per frame).

6. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and D. Plumbley, "Detection and classification of audio scenes and events," *IEEE Transaction on Multimedia*, vol. 17, no. 10, pp. 1733-1746, 2015.

[2] J. Rennies, S. Goetze, and J. .-E. Appell, "Personalized Acoustic Interfaces for Human-Computer Interaction," in *Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications*. IGI Global, 2011, ch. 8, pp. 180-207.

[3] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Speech and Multimodal Interaction in Assistive Environments*, Jeju Island, 2012.

[4] D. Giannoulis, et al., "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, 2013.

[5] J. Schröder, et al., "On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events'," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, 2013.

[6] R. Stiefelhagen, et al., "The clear 2006 evaluation," in *Multimodal technologies for perception of humans*. Springer Berlin Heidelberg, 2007, pp. 1-44.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference*, Budapest, 2016.

[8] A. Rakotomamonjy and G. Gasso, "Historgram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transaction on Audio, Speech, and Signal Processing*, vol. 23, no. 1, pp. 142-153, 2015.

[9] J. .-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-grames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881-891, 2007.

[10] W. Nogueira, G. Roma, and P. Herrera, "Sound scene identification based on MFCC, binaural features and a support vector machine classifier," technical report, 2013.

[11] G. Roma, W. Nogueira, and P. Herrera, "Recurrence quantification analysis features for audiotory scene classification," technical report, 2013.

[12] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognizing acoustic scenes with large-scale audio feature extraction and SVM," TUM, technical report, 2013.

[13] D. Li, J. Tam, and D. Toub, "Auditory scene classification using machine learning techniques," technical report, 2013.

[14] J. D. Krijnders and G. A. ten Holt, "A tone-fit feature representation for scene classification," technical report, 2013.

[15] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal Gabor filterbank features for acoustic event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 2198-2208, 2016.

[16] M. Chum, A. Habshush, A. Rahman, and C. Sang, "IEEE AASP scene classification challenge using hidden Markov models and frame based classification," technical report, 2013.

[17] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classificantion," in *23rd European Signal Processing Conference*, Nice, 2015, pp. 2551-2555.

[18] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classfication," M.S. thesis, ATIAM ParisTech, Paris, 2011.

[19] V. Bisot, R. Sterizal, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, 2016, pp. 6445-6449.

[20] N. Moritz, J. Anemüller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926-1937, 2015.

[21] N. Moritz, et al., "A CHiME-3 challenge system: Long-term acoustic features for noise robust automatic speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Phoenix, 2015.

[22] N. Moritz, B. Kollmeier, and J. Anemüller, "Integration of optimized modulation filter sets into deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[23] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[24] M. Seltzer, Y. Dong, and Y. Wang, "An investiagtion of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, 2013, pp. 7398-7402.

[25] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *22nd European Signal Processing Conference*, Lisbon, 2014, pp. 506-510.

[26] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Multi-label vs. combined single-label sound event detection with deep neural networks," in *23rd European Signal Processing Conference*, Nice, 2015, pp. 2551-2555.

[27] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech*, Florence, 2011, pp. 437-440.

[28] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transaction on Acoustics, Speech, and Language Processing*, vol. 37, no. 3, pp. 328-339, 1989.

[29] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of a long temporal contexts," in *Interspeech*, Dresden, 2015, pp. 2440-2444.

[30] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, vol. 19, Vancouver, 2007, pp. 153-160.