PERFORMANCE COMPARISON OF GMM, HMM AND DNN BASED APPROACHES FOR ACOUSTIC EVENT DETECTION WITHIN TASK 3 OF THE DCASE 2016 CHALLENGE

Jens Schröder[1,3*], *Jörn Anemüller*[2,3], *Stefan Goetze*[1,3]

[1] Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany
[2] University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[3] Cluster of Excellence, Hearing4all, Germany
jens.schroeder@idmt.fraunhofer.de

ABSTRACT

This contribution reports on the performance of systems for polyphonic acoustic event detection (AED) compared within the framework of the “detection and classification of acoustic scenes and events 2016” (DCASE’16) challenge. State-of-the-art Gaussian mixture model (GMM) and GMM-hidden Markov model (HMM) approaches are applied using Mel-frequency cepstral coefficients (MFCCs) and Gabor filterbank (GFB) features and a non-negative matrix factorization (NMF) based system. Furthermore, tandem and hybrid deep neural network (DNN)-HMM systems are adopted. All HMM systems that usually are of single label type, i.e., systems that only output one label per time segment from a set of possible classes, are extended to multi label classification systems that are a compound of single binary classifiers classifying between target and non-target classes and, thus, are capable of multi labeling. These systems are evaluated for the data of residential areas of Task 3 from the DCASE’16 challenge. It is shown that the DNN based system performs worse than the traditional systems for this task. Best results are achieved using GFB features in combination with a single label GMM-HMM approach.

Index Terms— acoustic event detection, DCASE’16, Gabor filterbank, deep neural network

1. INTRODUCTION

Acoustic event detection (AED) denotes the automatic identification of sound events in audio signals. Commonly, the acoustic event’s category as well as its time of occurrence are to be recognized. Application fields for AED are e.g., surveillance of public spaces for security issues [1–3], monitoring of health states e.g. in care systems [4–6] or condition monitoring of technical systems [7, 8].

AED in monophonic environments, i.e., for settings in which only single, isolated acoustic sources are active for a given time interval, has been the main focus of research in the past, with prominent comparisons of competitive systems in, e.g., the “classification of events, activities and relationships” (CLEAR’07) and “detection and classification of acoustic scenes and events 2013” (DCASE’13) challenges. Established methods for detecting acoustic events in monophonic environments are often based on Mel-frequency cepstral coefficient (MFCC) features and hidden Markov models (HMMs) using Gaussian mixture models (GMMs) as observation probability functions (GMM-HMM) [9–11]. These systems are denoted as single label classification systems since for a certain time segment they select one and only one label from a set of pre-trained classes based on maximum likelihood criteria or comparable scores. However, in many realistic environments rarely only a single source is active per time instance. Instead, usually multiple sources emit sound waves simultaneously leading to a mixed sound signal at a receiver. This case of multiple and overlapping sound signals is commonly referred to as polyphony. For acoustic event detection systems this case is by far more challenging than the monophonic case, not only because of the pure signal mixture of an unknown number of acoustic events present in the signal but also because training and test data can be considerably different due to the vast number of possibilities of event mixtures. Recently, polyphonic acoustic event detection has gained considerable attention, e.g., by being addressed in the DCASE’16 challenge. Some approaches for polyphonic event detection are based on MFCC and GMM-HMM classifiers. Using these back-ends, either binary classification between target events and universal background model is performed [12] or classification on multiple streams separated by non-negative matrix factorization (NMF) is conducted [11]. Further approaches apply NMF as part of feature extraction by thresholding the activations of the source code book [13, 14]. In recent publications, deep neural networks (DNNs) are used [3, 15, 16]. The output of the DNNs replaces the NMF-features, while the classification continues to rely on thresholded feature values. In the field of automatic speech recognition (ASR), DNNs are well-established and constitute the state-of-the-art baseline. Incorporation into recognition systems is based on two paradigms, tandem and hybrid approaches [17]. For the tandem approach, DNN features replace the MFCC features while the back-end is a conventional GMM-HMM classifier. Commonly, bottleneck features are used, for which one layer of the DNN acts as “bottleneck” with only a small number of neurons compared to the preceding and subsequent layers [17]. The hybrid approach uses DNNs as observation functions replacing the GMMs leading to DNN-HMM back-ends. They can be used with any kind of features. A common observation with DNNs is that they need more training data than for example GMM-HMM systems with MFCCs features.

This paper describes the authors contribution to the DCASE’16 challenge. It focuses on the subtask of Task 3 containing acoustic data recorded in residential environments (cf. Section 2). We investigate the performance of GMM-HMM systems using MFCCs and Gabor filterbank (GFB) features, the best scoring system of the DCASE’13 challenge, as well as NMF. Furthermore, we examine

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

Table 1: Event statistics of Task 3 for the residential area. Given are the number of events ('num. ev.'), the average duration ('av. dur.') and the total duration ('tot. dur.') of each class individually and overall as mean and standard deviation.

	num. ev.	av. dur. [s]	tot. dur. [s]
bird singing	130	7.55 ± 25.19	981.21
car passing by	57	9.16 ± 4.81	521.94
children shouting	23	2.00 ± 1.68	46.16
object banging	15	0.76 ± 0.70	11.33
people speaking	40	8.08 ± 24.42	323.08
people walking	32	5.50 ± 5.94	176.11
wind blowing	22	6.09 ± 5.98	133.96
overall	319	6.88 ± 18.59	2193.79

the performance of DNN tandem and hybrid approaches. Single label and multi label classification systems are used.

The remaining of this paper is structured as follows. The experimental setup including the dataset 'residential area' of Task 3 from the DCASE'16 challenge is outlined in Section 2. The concept of single label and multi label systems is explained in Section 3. The individual classification systems are detailed in Section 4. The results for these systems are shown in Section 5. Conclusions are drawn in Section 6.

2. EXPERIMENTAL SETUP

The following experiments are based on the setup and data of Task 3 of the DCASE'16 challenge [18]. Task 3, called 'Sound event detection in real life audio', consists of stereo data recorded at 44.1 kHz and in a home environment and in a residential area. Only the first channel is used in our contribution. The dataset of the home environment comprises eleven classes of a total duration of 36 min whilst the dataset of the residential area is a compound of seven classes and a total duration of 42 min. Since these are relatively few data especially for training of DNNs, we will just show results for the larger subset 'residential area'. Details of this subset are given in Table 1. The proposed four cross-validation sets from the challenge are used as well as the evaluation measures F-Score and the acoustic event error rate (AEER) [18]. The F-Score F represents the relation between the precision P and the recall R, i.e.,

$$P = \frac{N_{\text{corr}}}{N_{\text{est}}}; \quad R = \frac{N_{\text{corr}}}{N_{\text{ref}}}; \quad F = \frac{2 \cdot P \cdot R}{P + R}, \tag{1}$$

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_{ref}, i.e.

$$\text{AEER} = \frac{I + D + S}{N_{\text{ref}}}. \tag{2}$$

Both measures are applied on 1 sec segments and averaged over all crossvalidation folds.

3. SINGLE AND MULTI LABEL SYSTEMS

For detecting events, two main classification systems will be tested: Single label classification and multi label classification systems. A single label classification system consists of multiple models for different classes. The model yielding highest probability for a time segment is selected as label. Thus, such approaches are not capable of detecting simultaneous or overlapping events. Commonly, HMM systems are single label classification systems. To overcome this disadvantage and get multiple labels per time segment, the single label systems can be extended to multi label classification systems. A single binary classifier consists of a target class model and a garbage or background model that covers all non-target classes. Hence, a compound of such binary classifiers in a classification system is able to label each time segment with multiple labels.

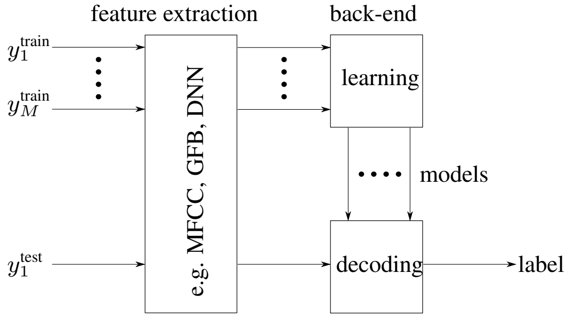

Figure 1: General schematic of the applied classification systems.

4. CLASSIFICATION SYSTEMS

The commonly applied classification systems consist of a feature extraction step and a back-end (cf. Figure 1). In the training phase, the extracted features, e.g. MFCCs, GFB features, DNN features etc., are used to create class models for the back-end that can be, e.g., HMMs. In the testing phase, these models are applied to the extracted features of the test data to decode it and output labels for time segments. The adopted systems of this contribution will be detailed in the following.

4.1. Baseline System

As baseline we use the provided baseline system from Task 3 of the DCASE'16 challenge [18]. It is composed of a GMM model using MFCCs. The MFCC features use 40 ms windows with 50% shift. The first 19 coefficients and the 0th energy coefficient plus derivations of first (Δ) and second order ($\Delta\Delta$) are used, that are computed over 9 time frames. The GMM is based on 16 Gaussian mixtures per class model and is applied on sliding windows of 1 second. The baseline is just applied in a binary classification system.

4.2. NMF System

The NMF system is based on the baseline system of Task 2 of the DCASE'16 challenge. It uses variable Q-transform (VQT) spectrograms of 60 bins per octave and a step size of 10 ms. The NMF codebook consists of 20 spectral templates per class that are learned during a training phase. For the original baseline, the 20 spectral templates were generated by averaging the delivered 20 event files

Table 2: Results of Task 3 for the residential area. In each row, the performance of the respective system is given in terms of AEER and F-Score. Both measures are divided into the total average and the class-wise average. A check mark in column ‘multi label’ indicates that the system is capable of making multi label outputs, e.g. the binary systems, otherwise systems produce single label output. For DNN features, the underlying features are given in brackets. Note: The baseline system uses other parameters for MFCCs than the other MFCC based systems depicted in rows 3, 4, 7, 8, 10 and 11. Best scores are highlighted by bold numbers.

no.	multi-label	back-end	feature	AEER		F-Score [%]	
				total	class	total	class
1	√	baseline	MFCC	0.86	1.16	34.6	19.9
2	√	NMF	VQT	1.35	2.11	14.8	8.7
3		GMM-HMM	MFCC	0.77	1.02	41.2	15.8
4	√	GMM-HMM	MFCC	0.81	1.08	41.0	16.6
5		DNN-HMM	log-Mel	1.02	3.47	17.2	8.5
6	√	DNN-HMM	log-Mel	1.78	5.87	16.0	13.3
7		DNN-HMM	MFCC	1.04	3.75	10.7	7.9
8	√	DNN-HMM	MFCC	1.22	2.86	22.6	14.7
9	√	GMM-HMM	DNN(log-Mel)	2.17	6.23	28.6	19.6
10	√	GMM-HMM	DNN(MFCC)	1.87	6.08	30.8	18.8
11	√	GMM-HMM	MFCC+DNN(log-Mel)	2.37	5.89	32.4	**24.1**
12		GMM-HMM	GFB	**0.74**	**1.01**	**48.5**	19.2
13	√	GMM-HMM	GFB	0.93	1.44	44.2	17.6

per class. Hence, the codebook size depended on the amount of files. To avoid the dependency on the dataset size, we modified the training phase by applying a GMM with 20 mixture components to the complete spectrogram data of each class to create the desired number of spectral templates. Based on these templates, data is decoded by a NMF. The NMF output is postprocessed using a threshold (1.0), a minimum event length of 60 ms and a maximum number of concurrent events (5).

4.3. DNN-HMM Hybrid System

For the DNN-HMM hybrid system, the commonly applied GMM observation function of an HMM is replaced by a DNN. The HMM for each class is modeled by one transition state, i.e., it is actually a GMM. Viterbi-decoding is applied with multiple, unlimited number of repetitions of events per file to get time segment labels. The input layer consists of the current time frame plus 4 frames before and after, thus, extending the feature dimensionality by a factor of 9. Several different combinations of number of layers (2,3,4), number of neurons per layer (20, 32, 39, 64, 128, 256) and characteristics like a bottleneck have been tested. Here, only the results of the DNN yielding best performance using three hidden layers with 128, 20, and 39 neurons will be shown. The hidden layers use the rectified linear unit (ReLU) as activation function, whilst the output function applies the softmax function.

Two types of features are investigated. One feature type is based on static MFCCs, i.e., a window length of 25 ms and 10 ms shift is used to compute the twelve first coefficients as well as the 0th. The other feature type is a logarithmic Mel (log-Mel)-spectrogam with 40 frequency bins (window length of 25 ms and shift of 10 ms).

4.4. GMM-HMM Sytem

The GMM-HMM systems use GMMs as observation functions for HMMs. The HMM of each class is modeled by one transition state. The best number of mixtures is evaluated on the validation fold, i.e., the performance of the mixture yielding the best total performance will be shown. In contrast to the baseline, the decoding is done using Viterbi-decoding with multiple repetitions of events per file.

Several different features are used for this system. Basic MFCCs as for the DNN-HMM hybrid system (cf. Section 4.3) with additional Δ and $\Delta\Delta$ features. Another feature type is based on the GFB. The GMM-HMM(GFB) system [19] achieved highest performance on the previous DCASE’13 challenge [20]. Here we use the GFB optimized for AED that has been shown to improve the results for the acdcase2013 challenge [21].

Furthermore, features are derived from DNNs, thus building a tandem system. Therefore, the DNNs of the hybrid systems are applied, and, hence, are either based on MFCCs or on the log-Mel-spectrogram. The hybrid DNNs are modified by deleting the output layer that represents the class probabilities, and replacing it by the second last layer containing 39 neurons. Furthermore, the activation function ReLU is replaced by a linear activation function to produce features with better discriminative abilities [22]. We used HTK [22] to adapt HMMs and DNNs.

5. RESULTS

The results of the tested systems are given in Table 2. The AEER and the F-Score are shown. They are divided into a total average over all frames and into a class-wise average, i.e., the score for each class is computed and the average of these numbers are depicted. Hence, effects on scores resulting from different amount of data per class are avoided. Each row describes a system. A check mark (√) in column ‘multi label’ indicates that a system has multi label output, which are the baseline system, the NMF system and the multi label versions of the HMM approaches. No check mark indicates that a system has single label output, which are the standard HMM versions.

It can be seen that the NMF based system (cf. Row 2), which is the baseline system of Task 2 of the DCASE’16 challenge, performs relatively poorly compared to the GMM(MFCC) baseline (Row 1). This might result from the polyphonic training data. Commonly, the training data for NMF approaches consist of isolated events.

However, the data for Task 3 was polyphonic. Thus, a proper codebook is unlikely to be generated leading to much confusion between classes. Another reason for the inaccuracy of the approach might be that beeing the baseline system of Task 2, the used parameters for the classifier may not be optimal for Task 3.

The GMM-HMM-systems using MFCCs (cf. Rows 3 and 4) achieve better results for the AEER and for the total F-Score. For the class-wise F-Score, they are slightly worse. This is a result of the unequal amount of data per class. Both GMM-HMM-systems are particularly good in recognition of the classes with most data 'bird singing' and 'car passing by'. For class 'bird singing', the single label GMM-HMM-systems (cf. Row 3) yields a class-wise F-Score of 56.3% whereas the baseline yields F-Score of 35.5%. This imbalance leads to a better total F-Score for the GMM-HMM-systems but a worse class-wise F-Score than for the baseline.

Against expectation, the single label approach of the GMM-HMM-system yields better performance than the multi label approach, though the single label approach is not capable of detecting multiple overlapping events and, thus, in contrast to the binary approach, by its nature can never yield 100% accuracy. For the applied dataset, it seems to be beneficial to just output one label with maximum likelihood than to try to detect multiple concurrent events.

However, for the DNN-HMM hybrid systems (cf. Rows 5 to 8), the binary versions yield better F-Scores than the single label systems. In comparison to the baseline, they perform worse for all shown measures. The tandem systems (cf. Rows 9 to 11) yield worse AEER scores. However, the F-Scores are relatively high. For the system with concatenated MFCC and DNN features (cf. Row 11), even the highest class-wise F-Score of all examined systems is achieved. The reason for the low AEER but high F-Score lies in the high number of label outputs that are generated by the tandem system. It causes many errors but also a high recall R forcing a relative high F-Score.

The best scores except for the class-wise F-Score are achieved by the single label GMM-HMM with GFB features (cf. Row 12). Especially the total F-Score is much higher than for all other tested systems. Similar to the GMM-HMM-systems using MFCCs (cf. Row 4), the multi label version of the GMM-HMM system adopting GFB features (cf. Row 13) performs less well than the single label version.

6. CONCLUSIONS

This study reports system performances for different acoustic event detection strategies applied to Task 3 ('residential area' data) of the DCASE'16 challenge. We compared commonly used GMM systems to tandem and hybrid DNN systems. Single and multi label systems were applied. We showed that for this task, DNNs are less accurate than GMM-HMM systems. Probably, the amount of data available for Task 3 is too low to train DNNs properly. The GMM-HMM system in combination with GFB features which was developed for the DCASE'13 challenge [19] for isolated events and that was meanwhile improved as described in [21], yields best performance of all tested systems. Furthermore, a single label system that is only able to output one label per time segment seems not to be inferior to a multi label classification system for polyphonic data.

A drawback for all classification systems is the little amount of available data within this challenge. As it could be observed, the two classes with most data, which are 'bird singing' and 'car passing by', achieved best recognition results for nearly all classification systems. Testing the approaches on larger corpora is thus subject of future work.

7. REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, Jul. 2005, pp. 1306–1309.

[2] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic acoustic siren 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, pp. 493–497.

[3] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6460–6464.

[4] S. Päßler and W. Fischer, "Food intake monitoring: Automated chew event detection in chewing sounds," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 278–289, 2014.

[5] J. Schröder, J. Anemüller, and S. Goetze, "Classification of human cough signals using spectro-temporal Gabor filterbank features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6455–6459.

[6] S. Matos, S. S. Birring, I. D. Pavord, and D. H. Evans, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1078–1083, 2006.

[7] J. Schröder, M. Brandes, D. Hollosi, J. Wellmann, M. Wittorf, O. Jung, V. Grützmacher, and S. Goetze, "Foreign object detection in tires by acoustic event detection," in *DAGA 2015*, Nuremberg, Germany, Mar. 2015, pp. 1266–1269.

[8] N. K. Verma, R. K. Sevakula, S. Dixit, and A. Salour, "Intelligent condition based monitoring using acoustic signals for air compressors," *IEEE Trans. Reliability*, vol. 65, no. 1, pp. 291–309, 2016.

[9] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[10] 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.

[11] T. Heittola, A. Mesaros, A. J. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal of Audio, Speech and Music Processing*, vol. 2013, p. 1, 2013.

[12] J. Schröder, F. X. Nsabimana, J. Rennies, D. Hollosi, and S. Goetze, "Automatic detection of relevant acoustic events in kindergarten noisy environments," in *DAGA 2015*, Nuremberg, Germany, Mar. 2015, pp. 1525–1528.

[13] O. Dikmen and A. Mesaros, “Sound event detection using non-negative dictionaries learned from annotated overlapping events,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[14] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, “Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 151–155.

[15] 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, EUSIPCO 2015*, Nice, France, Aug. 2015, pp. 2551–2555.

[16] A. Diment, E. Cakir, T. Heittola, and T. Virtanen, “Automatic recognition of environmental sound events using all-pole group delay features,” in *23rd European Signal Processing Conference, EUSIPCO*, Nice, France, Aug. 2015, pp. 729–733.

[17] Z. Tüske, R. Schlüter, H. Ney, and M. Sundermeyer, “Context-dependent MLPs for LVCSR: TANDEM, hybrid or both?” in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, Sep. 2012, pp. 18–21.

[18] A. Mesaros, T. Heittola, , and T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” in *24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, Sep. 2016, p. ?, *accepted*.

[19] J. Schröder, N. Moritz, M. R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, “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 (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[20] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: An IEEE AASP challenge,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[21] 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, no. 12, pp. 2198–2208, Dec. 2015.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. R. V. Valtchev, P. Woodland, and C. Zhang, *The HTK Book (for HTK Version 3.5alpha)*, 2015.