A BAG-OF-FEATURES APPROACH TO ACOUSTIC EVENT DETECTION

Axel Plinge, René Grzeszick, and Gernot A. Fink

Department of Computer Science, TU Dortmund University, Dortmund, Germany

ABSTRACT

The classification of acoustic events in indoor environments is an important task for many practical applications in smart environments. In this paper a novel approach for classifying acoustic events that is based on a Bag-of-Features approach is proposed. Mel and gammatone frequency cepstral coefficients that originate from psychoacoustic models are used as input features for the Bag-of representation. Rather than using a prior classification or segmentation step to eliminate silence and background noise, Bag-of-Features representations are learned for a background class. Supervised learning of codebooks and temporal coding are shown to improve the recognition rates. Three different databases are used for the experiments: the CLEAR sound event dataset, the D-CASE event dataset and a new set of smart room recordings.

Index Terms— Event detection, sound classification, Bag-of-Features

1. INTRODUCTION

The classification of sounds in indoor environments is important for many practical applications. The detection and classification of acoustic events can be used for meeting and online lecture analysis and annotation [1]. For speech enhancement and speaker tracking [2] detecting non-speech events can improve the robustness in real world applications.

The task is difficult because of the diversity of the acoustic events. Human speech is comprised of sounds of different phone classes, e.g. vowels, plosives and fricatives that have individual spectrum and time characteristics. Other sound types are also complex because they are comprised of a variety of individual sounds, e.g. chair movement can produce knocking and rubbing sounds, handling paper can include rustling and knocking on the table and so on. Sounds like footsteps are individually different depending on the person and kind of shoes. It is desirable for a sound classification method to be able to handle the diverse composition and generalize in a way to cover different, possibly unheard realizations of the sound types.

Over the last decades, a number of approaches for acoustic event detection have been proposed [3–5]. State-of-the-art approaches in speaker classification are based on a single Gaussian mixture model (GMM) [6]. Others use a set of GMMs that are individually trained for each class, where the GMMs estimates are summed over all frames and the class with the highest likelihood is chosen. Since the summation discards any temporal information, the method is sometimes termed 'Bag-of-Frames' [5,7]. Since considerable progress has been made by applying insights from human perception in the field of computer or machine vision, similar approaches have been advocated for acoustics [8].

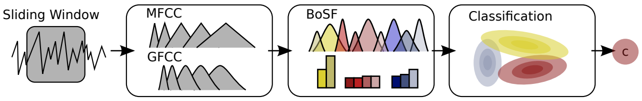

Fig. 1. Overview of the Bag-of-Features based method that is used for acoustic event detection.

The Bag-of-Features approach originated in text retrieval [9]. It has successfully been used in various pattern recognition applications in recent years [10–12]. For example, in image classification the Bag-of-Features is known to generalize well over very diverse classes, producing state-of-the-art results [11]. Recently, a basic version was applied to acoustic event classification [13].

In this paper a novel Bag-of-Features approach based on soft quatization with GMMs is introduced. Experiments show that is able to distinguish very diverse sound event classes.

2. METHOD

For the acoustic event detection and classification, a single microphone or beamformed signal is processed in short time windows of $0.6\,\mathrm{s}$ every $0.05\,\mathrm{s}$. For a given sample n, a set of feature vectors $Y_n = (y_1 \ldots y_K)$ is calculated for all frames in this window. These features are then softly quantized by a GMM and classified by an multinomial maximum likelihood classifier. Rather than using a prior classification step to eliminate silence and background noise, as done in several systems (cf. [3]), the rejection class Ω_0 is trained with recordings where no event occurred.

2.1. Features

For sound and especially speech processing, the mel frequency cepstral coefficients (MFCCs) are one of the most

This work was supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

widely used features. The input signal is filtered by a mel frequency filter bank, from the logarithm of its magnitude the discrete cosine transform (DCT) is computed and its second to 13th coefficient is used.

The long history of psychoacoustic research has been complemented by computational modeling of the human hearing process [14] where ERB-spaced gammatone filterbanks are used. From that the gammatone frequency cepstral coefficients (GFCCs) were derived [15]. In our implementation, we replaced the filterbank of the MFCCs by linear phase gammatone filters. The filters are defined in the spectral domain using a gammatone approximation [16] with center frequency f_b and bandwidth w_b

$$G^{(b)}(f) = \left(1 + \mathrm{j}(f - f_b)/w_b\right)^{-4}, \tag{1}$$

where j is the imaginary unit.

2.2. Bag-of-Super-Features

A Bag-of-Features approach (cf. [17]) is used for building a codebook of *acoustic words* from the training set. Most Bag-of-Features approaches use clustering algorithms, e.g. the Lloyd algorithm, on the complete training set to derive a codebook and later assign each feature to a centroid by hard quantization. However, disregarding the labels in the clustering step can lead to mitigation of significant differences (cf. [18]). A remedy for this effect is to build codebooks of size I for all C classes Ω_c separately and then concatenating into a large super-codebook . Here, the expectation-maximization (EM) algorithm is applied to all feature vectors $\boldsymbol{y}_k$ for each class Ω_c in order to estimate I means and deviations $\mu_{i,c}, \sigma_{i,c}$ for all C classes. We concatenate all means and deviations into a super-codebook $\boldsymbol{v}$ with $L = I \cdot C$ elements

$$v_{l=(I \cdot c+i)} = (\mu_{i,c}, \sigma_{i,c}) \tag{2}$$

where the index l computed form the class index c and the Gaussian index i as $l = I \cdot c + i$. Using this codebook, a soft quantization of a feature vector $\boldsymbol{y}_k$ can be computed as

$$q_{k,l}(\boldsymbol{y}_k, v_l) = \mathcal{N}(\boldsymbol{y}_k | \mu_l, \sigma_l) \,. \tag{3}$$

Then, a histogram $\boldsymbol{b}$ can be computed over all K frames of the input window by

$$b_l(Y_n, v_l) = \frac{1}{K} \sum_k q_{k,l}(\boldsymbol{y}_k, v_l) \,. \tag{4}$$

We refer to this method as ”Bag-of-Super-Features” in analogy to the super-vector construct used in speaker identification [6].

2.3. Temportal Pyramid

Since a Bag-of-Features is an orderless representation all temporal information within the frame Y_n is lost. However, this information may be important for distinguishing different acoustic events. In the last years several approaches have been published in order to address this problem. For example, spatial features [12] or pyramids [19].

The pyramid scheme is directly applied to the Bag-of-Super-Features approach by subdividing the window in a temporal manner. For a feature vector of the n^{th} window two sub-histograms

$$b_l^{(1)}(Y_n, v_l) = \frac{2}{K} \sum_{k=1}^{K/2} q_{k,l}(\boldsymbol{y}_k, v_l) \quad \text{and}$$
$$b_l^{(2)}(Y_n, v_l) = \frac{2}{K} \sum_{k=K/2+1}^{K} q_{k,l}(\boldsymbol{y}_k, v_l) \tag{5}$$

are defined for the first and the second temporal half. In addition, a max pooling step is used for computing the histogram for the whole window by

$$b_l^{(3)}(Y_n, v_l) = \max\left\{ b_l^{(1)}(Y_n, v_l), b_l^{(2)}(Y_n, v_l) \right\} \,. \tag{6}$$

All three histograms are then concatenated into a single feature vector

$$\boldsymbol{b}(Y_n, \boldsymbol{v}) = \left(\boldsymbol{b}^{(1)}(Y_n, \boldsymbol{v}), \boldsymbol{b}^{(2)}(Y_n, \boldsymbol{v}), \boldsymbol{b}^{(3)}(Y_n, \boldsymbol{v}) \right) \tag{7}$$

that represents the complete window.

2.4. Classification

The probability of an acoustic word for a given class $P(v_l|\Omega_c)$ is estimated using a set of training samples $Y_n \in \Omega_c$ for each class c by Laplacian smoothing:

$$P(v_l|\Omega_c) = \frac{1 + \sum_{Y_n \in \Omega_c} b_l(Y_n, v_l)}{L + \sum_{m=1}^{L} \sum_{Y_n \in \Omega_c} b_m(Y_n, v_m)} \tag{8}$$

Since all classes are assumed to be equally likely and have the same prior, maximum likelihood classification is used. The posterior is estimated using the relative frequency of all acoustic words

$$P(Y_n|\Omega_c) = \prod_{v_l \in \boldsymbol{v}} P(v_l|\Omega_c)^{b_l(Y_n, v_l)} \,. \tag{9}$$

3. EVALUATION

In order to derive a working system for the smart room at TU Dortmund University, several recordings were made. Different features were evaluated using the proposed classification method. The proposed method and related ones were compared in classification performance. The event detection capability was tested with a scripted recording in the smart room and several others from existing corpora.

3.1. Event Classification & Model Parameters

In order to investigate the performance of different methods, recordings of various typical sound events were made in a smart conference room at TU Dortmund University. The microphones were embedded in a table as shown in Fig. 2 and recorded at 48 kHz. Each recording featured a certain sound type and lasted over 60 s.

To evaluate the classification performance on unknown data, a second test set of recordings was made on a different day with a different person producing the sounds. In the recordings time stretches with occurrences of the events were labeled. All methods were evaluated using cross-validation on the training and test set.

Using the Bag-of-Super-Features-Pyramid (BoSF-P) approach, different feature types were evaluated. Figure 3 shows the results. Along with the MFCCs the GFCCs, linear prediction coefficients (LPC) and a non-negative matrix factorization (NMF) [20] of the mel frequnecy magnitudes were evaluated. The MFCCs and GFFCs have the lowest error on the test set. Their combination achieves the highest score. Both LPC and NMF show a significantly higher error on the test set and seem to be unable to generalize successfully.

Figure 4 shows the classification errors for different methods using a combination of MFCC and GFCC features. The Bag-of-Frames model (BoFr) using MFCCs only that is described in [5] is applied to the sound classification problem and used as a baseline. The Bag-of-features (BoF) model performs worse than the baseline if the codebook is computed in an unsupervised manner. However, there is a significant improvement using the Bag-of-Super-Features (BoSF). This strengthens the view that the use of a supervised codebook estimation allows for a better modeling of the diverse acoustic event classes. Incorporating temporal information by the pyramid scheme further improves the results. For comparison, a Nearest Neighbor classifier and an SVM were also applied to the pyramid model. They both perform significantly worse than the multinomial maximum likelihood classifier.

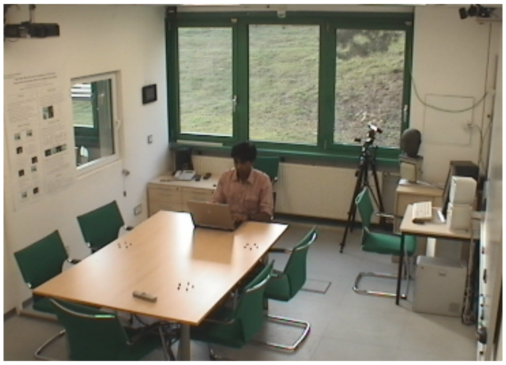

Fig. 2. Smart room with microphones embedded in table.

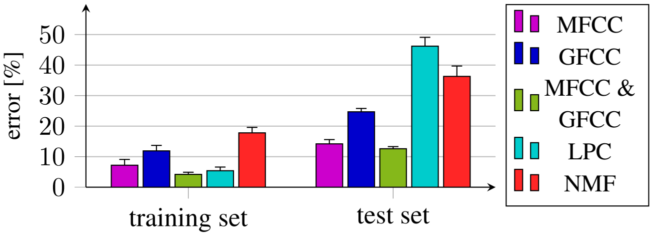

Fig. 3. Classification error for different features of smart room recordings. All features were evaluation with the BoFS-P method using a multinomial maximum likelihood classifier.

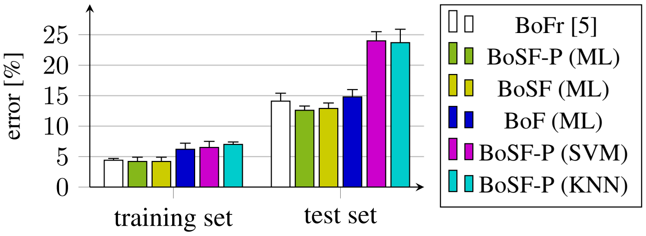

Fig. 4. Classification error for different methods for the smart room recordings. The Bag-of-Feature methods were evaluated using the combination of MFCC and GFCC features.

In order to determine the influence of the codebook size the BoSF-P approach has been evaluated for different sizes of L. The results in Fig. 5 show that already a comparably small codebook size of $L = 121$ yields good results, which equals 11 centroids per class. Therefore, in the following experiments a codebook size of 11 centroids per class was chosen. Compared to other Bag-of-Features classification approaches where codebooks of several thousand centroids are used this size is remarkably small (see [10,11]). The advantage of this is two-fold: First, the quantization adds an additional abstraction to the data such that it generalizes better. Large codebooks approximate the data better but are not able to generalize well over the very diverse acoustic events. Second, small feature representations are fast to compute and classify which facilitates the use of the method in real time acoustic event detection. The proposed method can be computed in just 5% of the real time using python on a standard PC.

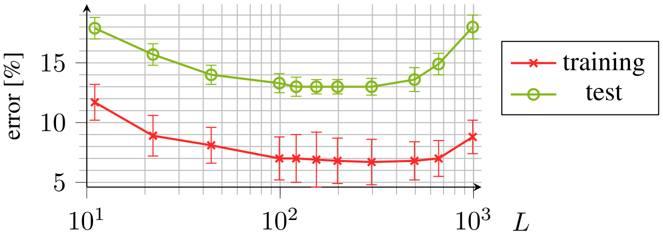

Fig. 5. Classification error for the BoSF-P approach for the smart room recordings using MFCC and GFCC features with different codebook sizes L.

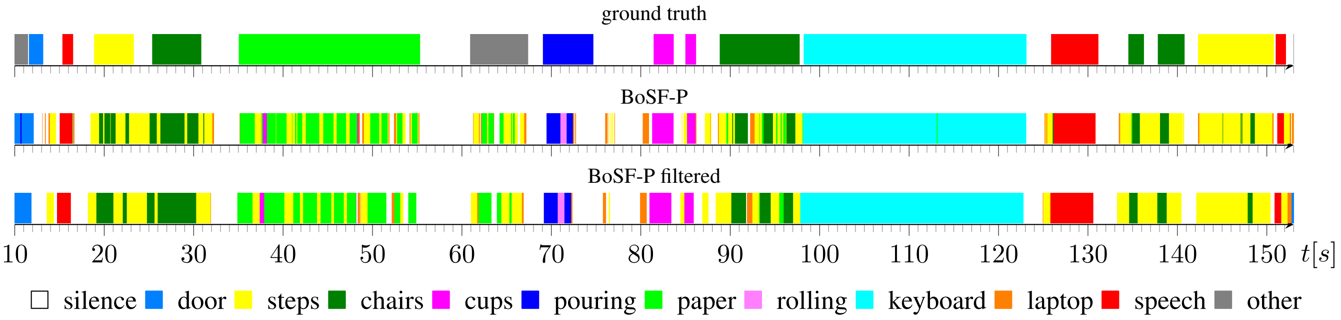

Fig. 6. Results for event detection in the smart environment, acoustic events are shown in distinct colors.

3.2. Event Detection

When denoting as g, e, and t the number of ground truth, estimated and correct events, precision P and recall R can be defined along with the F-measure F as in [5]

$$P = \frac{t}{e}\,, \quad R = \frac{t}{g}\,, \quad F = \frac{2PR}{P+R}\,. \tag{10}$$

For the event detection performance, the non-event class Ω_c is excluded in the counts. The metrics are evaluated frame-based and class-based, for the latter all classes are evaluated individually and the average is computed.

3.2.1. Smart Room Recording

In order to establish the systems performance for event detection in live scenarios, sequences with various events were recorded in the smart room. The classifier was chosen by the evaluations above and trained again with the event recordings. Table 1 lists the overall detection metrics. The non-event class had 83% precision and 68% recall. This can be attributed to the fact that the training data for other classes contained portions of silence. The 'speech' class was detected with 97% precision and 87% recall. In Fig. 6, the detection results for the sequence are visualized in color. Smoothing may be desirable for practical applications. Basic post filtering can be done by selecting the most frequent detection in the last 1 s and discarding cases where its occurrence covers less than 0.3 s.

3.2.2. CLEAR

Within the CHIL project, the CLEAR campaign investigated the detection of acoustic events. The proposed method was tested on the ITC data, which contains three different training sets and a test set for three separate days [3]. For the non-event class, the non-labeled portions from the training data were used. In this manner, 88% precision and 84% recall were achieved. Table 1 shows the performance over all experiments in the development set. The 'phone vibration' class had 0% recall, for all other classes an F value of over 75% was achieved.

dataset	method	metric	F	P	R
Smart Room	BoSF-P	frames	71.9%	74.3%	69.6%
		classes	77.3%	82.7%	72.5%
CLEAR	BoSF-P	frames	75.8%	79.3%	72.6%
		classes	75.5%	79.2%	72.2%
D-CASE	BoSF-P	frames	52.3%	51.7%	53.2%
		classes	59.5%	64.8%	57.7%
	NMF [5] baseline	frames	20.6%	29.1%	16.0%
		classes	13.5%	11.6%	21.7%

Table 1. Results for the acoustic event detection on the three datasets: smart room recordings, CLEAR and DCASE.

3.2.3. D-CASE

The proposed method was evaluated on the recent IEEE AASP Challenge 'Detection and Classification of Acoustic Scenes and Events' Event Detection development set. Since the training data consists of event recordings only, non-labeled portions of the scripts not used in the test were used for training in order to have training data for the non-event class. The performance averaged over all experiments in the development set are presented in table 1. The non-event detection had 88% precision and 90% recall. The baseline event detection proposed in [5] is outperformed and the results of our method are also highly competitive with respect to the results of the challenge.[1]

4. CONCLUSION

In this paper an event detection approach using supervised trained GMM codebooks of MFCCs an GFCCs for Bag-of-Features histograms with temporal coding was presented. Highly competitive results on different difficult datasets for acoustic event classification and detection were achieved. The use of a single 'silence' class for non-events could be shown to be highly successful. The good speech detection quality is important for many applications. The method can be easily implemented and computed fast enough for real-time application.

[1] Results are to be published (http://www.waspaa.com/d-case-challenge).

5. REFERENCES

[1] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic analysis of multimodal group actions in meetings.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 305–17, Mar. 2005.

[2] Axel Plinge, Daniel Hauschildt, Marius H Hennecke, and Gernot A Fink, "Multiple Speaker Tracking using a Microphone Array by Combining Auditory Processing and a Gaussian Mixture Cardinalized Probability Hypothesis Density Filter," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 2476–2479.

[3] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," in *Multimodal Technologies for Perception of Humans*, Rainer Stiefelhagen and John Garofolo, Eds., vol. 4122 of *Lecture Notes in Computer Science*, pp. 311–322. Springer Berlin Heidelberg, 2007.

[4] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen, "Acoustic Event Detection in Real-life Recordings," in *European Signal Processing Conference*, Aalborg, Denmark, 2010, pp. 1267–1271.

[5] Dimitrios Giannoulis, Dan Stowell, Emmanouil Benetos, Mathias Rossignol, and Mathieu Lagrange, "A Database and Challenge for Acoustic Scene Classification and Event Detection," in *European Signal Processing Conference*, Marrakech, Morocco, 2013.

[6] Hao Tang, Stephen M Chu, Mark Hasegawa-Johnson, and Thomas S Huang, "Partially supervised speaker clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 5, pp. 959–971, 2012.

[7] Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[8] Richard F. Lyon, "Machine Hearing – An Emerging Field," *IEEE Signal Processing Magazine*, Sept. 2010.

[9] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999.

[10] Leonard Rothacker, Marcal Rusinol, and Gernot A. Fink, "Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents," in *Proc. Int. Conf. on Document Analysis and Recognition*, Washington DC, USA, 2013.

[11] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011.

[12] René Grzeszick, Leonard Rothacker, and Gernot A. Fink, "Bag-of-features representations using spatial visual vocabularies for object classification," in *IEEE Intl. Conf. on Image Processing*, Melbourne, Australia, 2013.

[13] Stephanie Pancoast and Murat Akbacak, "Bag-of-Audio-Words Approach for Multimedia Event Classification," in *Interspeech*, Portland, OR, USA, 2012.

[14] DeLiang Wang and Guy J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE Press, 2006.

[15] Yang Shao, Soundararajan Srinivasan, and DeLiang Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 277–280.

[16] Masashi Unoki and Masato Akagi, "A Method of Signal Extraction from Noisy Signal based on Auditory Scene Analysis," *Speech Communication*, vol. 27, no. 3, pp. 261–279, 1999.

[17] Stephen O'Hara and Bruce A Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *arXiv preprint arXiv:1101.3354*, 2011.

[18] Svetlana Lazebnik and Maxim Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1294–1309, 2009.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[20] Daniel D. Lee and H. Sebastian Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization.," *Nature*, vol. 401, no. 6755, pp. 788–91, Oct. 1999.