Minimum Classification Error Based Spectro-Temporal Feature Extraction for Robust Audio Classification

Yuan-Fu Liao, Chia-Hsing Lin and We-Der Fang

Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan
yfliao@ntut.edu.tw

Abstract

Mel-frequency cepstral coefficients (MFCCs) are the most popular features for automatic audio classification (AAC). However, MFCCs are often not robust in adverse environment. In this paper, a minimum classification error (MCE)-based method is proposed to extract new and robust spectro-temporal features as alternatives to MFCCs. The robustness of the proposed new features is evaluated on noisy non-speech sound of RWCP Sound Scene Database in Real Acoustic Environment database with Aurora 2 multi-condition training task-like settings. Experimental results show the proposed new features achieved the lowest average recognition error rate of 3.17% which is much better than state-of-the-art MFCCs plus mean subtraction, variance normalization and ARMA filtering (MFCC+MVA, 4.31%), Gabor filters with principle component analysis (Gabor+PCA, 4.43%) and linear discriminant analysis (LDA, 4.20%) features. We thus confirm the robustness of the proposed spectro-temporal feature extraction approach.

Index Terms— spectro-temporal feature extraction, robust audio classification, minimum classification error

1. Introduction

MFCCs are currently the most popular features not only for ASR but also for AAC. However, MFCCs consider only local spectral information and ignore temporal cues. Therefore, MFCCs are often not so robust in noisy environment. Usually some extra feature normalization, model compensation methods or even hybrid system architecture, for example MVA [1], eigen-maximum likelihood linear regression (EMLLR) [2] or Tandem system [3], respectively, are necessary to alleviate the robustness issues.

Recently, spectro-temporal modulation selective filters, such as 2-dimensional Gabor filters, edge detectors and box filters are becoming popular for extracting new features as alternatives to MFCCs because they could simultaneously capture spectral and temporal modulation frequencies. This is aspired by findings in neuroscience which show neurons in the mammalian auditory cortex are highly tuned to specific spectro-temporal receptive fields (STRFs) [4].

For example, Gabor filters were used as a frontend to extract features for ASR in [5-6]. In [7] a set of time-frequency patches (edge detectors) was designed to detect some very specific phonetic events, such as formant transitions and phone boundaries. Gabor filters were also used in an auditory model for speech and non-speech discrimination [8]. Besides, in [9] six types of box filters were used to extract low-level features at multiple time scales for audio classification, recently.

One key issue to the success of these spectro-temporal filter-based approaches is how to select/merge a suitable set of filters/features since there is a large number of possible filter (1) type and (2) parameter combinations. In [5], Gabor filters were chosen by dividing the temporal modulation frequency from 1 to 16 Hz and the spectral modulation frequency from zero to two cycles per octave equally. In [6] a supervised trained feature-finding neural network (FFNN) was used to adjust the parameters of Gabor filters. In [7] Adaboost algorithm was applied to both build classifiers and perform feature selection from a large set of edge detectors. Besides, in [9] a GentleBoost algorithm was applied to construct a classifier that combines a subset of all possible box filters. Another popular solution is to apply PCA after a large set of arbitrary Gabor filters to reduce feature dimension [8].

In this paper, we focus on how to extract new and robust spectro-temporal features, as alternatives to MFCCs, in order to directly meet the final goal of audio classification, i.e., minimum classification errors. To this end, a MCE [10] algorithm is applied to simultaneously learn an optimal set of spectro-temporal filter type and parameter combinations (called feature extraction matrix from now on) from the feedback of an underlying audio recognizer. It is also worth noting that unlike those feature selection-based approaches in [5-9] which usually choose and fix the filter types in advance, no constraint is set on the feature extraction matrix (and extracted features) at all. In other words, the content of the feature extraction matrix may not necessary to be Gabor filters, edge detectors or box filters.

The paper is organized as follows. Section 2 briefly reviews the spectro-temporal feature extraction framework. In Section 3, a MCE-based approach is proposed to extract new features. Section 4 reports the audio classification results evaluated on noisy non-speech sound of RWCP Sound Scene Database in Real Acoustic Environment database [11] with Aurora 2 multi-condition training [12] task-like experiment settings. Some conclusions are drawn in the last section.

2. Spectro-Temporal Feature Extraction

The general framework of spectro-temporal feature extraction is briefly described here. It is worth noting that Gabor filters, edge detectors and box filters could be treated as a special case of this framework.

2.1. General framework

Fig. 1 shows a general framework of spectro-temporal feature extraction. This approach could be formulated as follows:

$$y_t = E^T x_t \tag{1}$$

where $x_t = \{o_{t-M}^T, ..., o_t^T, ..., o_{t+M}^T\}^T$ is the I-dimensional input super-vector in the t-th frame, which is made up of $2M+1$ neighboring raw filterbank feature vectors o_t, $E = \{e_{i,j}; i = 1 \sim I, j = 1 \sim J\}$ is the feature extraction matrix formed by a set of spectro-temporal filters, y_t is the new extracted J-dimensional feature vector.

It can be seen from Eq. (1), there could be a large number of possible filter type and parameter combinations for the feature extraction matrix E. Therefore, Gabor filters, edge detectors or box filters are often chosen in practice.

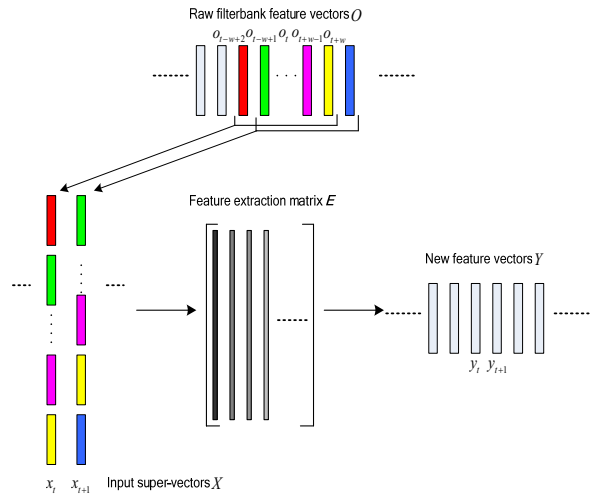

Fig. 1: A general block diagrams of the spectro-temporal feature extraction framework for audio classification.

3. The Proposed MCE-Based Approach

Fig. 2 shows a block diagram of the proposed MCE-based framework. As shown in Fig. 2, the MCE criterion could be used to optimize both (1) the frontend feature extraction matrix and (2) the backend hidden Markov models (HMMs).

However, in the following subsections, only optimization of the feature extraction matrix, E, will be discussed. The detail MCE-based algorithm for adjusting HMMs could be found in [10].

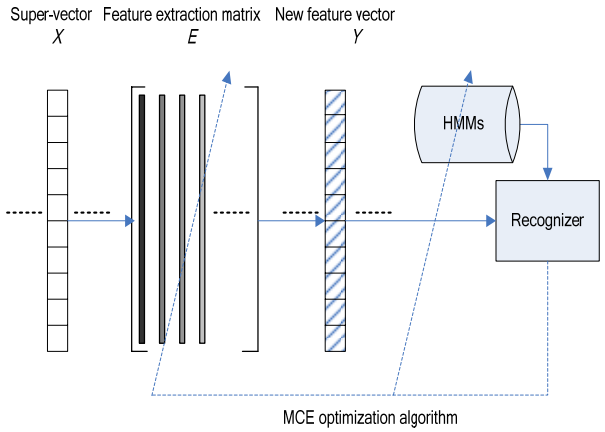

Fig. 2: Block diagram of the proposed MCE-based feature extraction framework for robust audio classification.

3.1. MCE criterion

Given an input test feature vector sequence Y, a score function $g_k(Y;\Lambda)$ is first defined for each audio class (K classes in total) using the log-likelihood functions $P_k(Y;\Lambda)$ of their corresponding HMMs Λ, i.e.,

$$g_k(Y;\Lambda) = \log\{P_k(Y;\Lambda)\}, k = 1 \sim K \tag{2}$$

Then, the decision rule of the recognizer is to choose the one with maximum score:

$$\hat{k} = \arg\max_k g_k(Y;\Lambda) \tag{3}$$

Next, if the input test sequence in fact belongs to the k'–th class, a miss-classification function $d(Y)$ is defined for the decision making:

$$d(Y) = -g_{k'}(Y;\Lambda) + \left(\frac{1}{K-1}\sum_{l\neq k'}\exp(\eta \cdot g_l(Y;\Lambda))\right)^{1/\eta} \tag{4}$$

where $g_l(Y;\Lambda)$ are the scores of competitive candidates other than k' and η is a constant to control the non-linearity of the miss-classification function .

If only the most competitive candidate, i.e., the k^* -th class, is considered, η is usually set to ∞, and the miss-classification function could be further simplified into Eq. (5).

$$d(Y) = -g_{k'}(Y;\Lambda) + g_{k^*}(Y;\Lambda) \tag{5}$$

Finally, a loss function $L(d(Y))$ is defined using a smooth "zero-one" Sigmod function:

$$L(d(Y)) = \frac{1}{1+e^{-(\gamma d(Y)+\theta)}} \tag{6}$$

where γ and θ are two constants that control the slope and bias of the Sigmod function.

It is worth noting that the summation of the loss functions over a set of test data approximates the empirical error counts of the test set.

3.2. GPD-based extraction matrix optimization

To properly adjust each element $e_{i,j}$ of the feature extraction matrix E, an iterative training algorithm using generalized probabilistic descent (GPD) [10] is applied here, i.e.,

$$e_{i,j}(n+1) = e_{i,j}(n) - \varepsilon_n \cdot \frac{\partial L(d(Y))}{\partial e_{i,j}} \tag{7}$$

where ε_n is a constant that control the learning step in the n -th iteration.

The last term of Eq. (7) could be further derived as follows:

$$\frac{\partial L(d(Y))}{\partial e_{i,j}} = \gamma \cdot L(d(Y)) \cdot (1 - L(d(Y))) \frac{\partial d(Y)}{\partial e_{i,j}}, \tag{8}$$

where

$$\frac{\partial d(Y)}{\partial e_{i,j}} = -\frac{\partial g_{k'}(Y;\Lambda)}{\partial e_{i,j}} + \frac{\partial g_{k^*}(Y;\Lambda)}{\partial e_{i,j}} \tag{9}$$

If partition Gaussian mixture is used in the HMMs, Λ, the score function $g_{k'}(Y;\Lambda)$ in Eq. (9) could be simplified as follows:

$$g_{k'}(Y;\Lambda) = \sum_{t=0}^{L-1}\log\left\{\omega^{k'}_{q_t,m_t} N(y_t;\mu^{k'}_{q_t,m_t},\sigma^{k'}_{q_t,m_t})\right\} \tag{10}$$

where m_t is the index of the largest mixture in state q_t in t –th frame, $N(y_t;\mu^{k'}_{q_t,m_t},\sigma^{k'}_{q_t,m_t})$ is the Gaussian function with $\omega^{k'}_{q_t,m_t},\mu^{k'}_{q_t,m_t},\sigma^{k'}_{q_t,m_t}$ are the corresponding mixture weight, mean and variance vectors.

And the partial differential equation, $\frac{\partial g_{k'}(Y;\Lambda)}{\partial e_{i,j}}$, becomes:

$$\begin{aligned}\frac{\partial g_{k'}(Y;\Lambda)}{\partial e_{i,j}} &= \sum_{t=0}^{L-1}\frac{\partial}{\partial e_{i,j}}\left[-\frac{1}{2}\sum_{j=1}^{J}\frac{(y_{t,j}-\mu^{k',m_t}_{q_t,j})^2}{\sigma^{k',m_t}_{q_t,j}}\right] \\ &= -\sum_{t=0}^{L-1}\frac{y_{t,j}-\mu^{k',m_t}_{q_t,j}}{\sigma^{k',m_t}_{q_t,j}}\cdot\frac{\partial y_{t,j}}{\partial e_{i,j}}\cdot x_{t,i}\end{aligned} \tag{11}$$

Therefore, the final formulation for $\frac{\partial L(d(Y))}{\partial e_{i,j}}$ is as follows:

$$\begin{aligned}&\frac{\partial L(Y;\Lambda)}{\partial e_{lj}} = \gamma\cdot L(d)\big(1-L(d)\big)\cdot \\ &\sum_{t=1}^{L-1}\left(\frac{y_{t,j}-\mu^{k',m_t}_{q_t,j}}{\sigma^{k',m_t}_{q_t,j}}-\frac{y_{t,j}-\mu^{k^*,m^*_t}_{q^*_t,j}}{\sigma^{k^*,m^*_t}_{q^*_t,j}}\right)\cdot x_{t,i}\end{aligned} \tag{12}$$

It is worth noting that Eq. (12) implies that the each element in the feature extraction matrix will be adjusted according to the product of the raw filterbank features and the difference vectors between the means of correct and most competitive mixtures. In other words, the most discriminative/robustness spectro-temporal cues will be automatically picked up by the MCE-based method.

4. Experimental Results

The robustness of proposed approach is evaluated and compared with several state-of-the-art methods on noisy non-speech sound of RWCP Sound Scene Database in Real Acoustic Environment database with Aurora 2 multi-condition training task-like experimental settings.

4.1. RWCP non-speech sound database

The recording of this database was conducted by RWCP in an anechoic room with 48 kHz sampling rate and 16 bits PCM data format. Each sound source was experimented by changing beating and ringing manners to yield many samples necessary for study on sound source discrimination.

There are in total 105 types of sound sources. Each sound source has about 100 samples. Table 1 shows a list of all sound sources. In all the following experiments, this database was down-sampled to 8 kHz and randomly divided into a training and a test set (60% and 40% samples of each sound type, respectively).

4.2. Noisy RWCP non-speech sound database

Following the same evaluation protocol used in Aurora 2 multi-condition training tasks, noises from Aurora 2 database and FaNT [13] toolkit were used to add artificial noise and communication channel characteristics to each sound sample.

There are eight noises in Aurora 2 database, including (1) subway, (2) babble, (3) car, (4) inhibition, (5) restaurant, (6) street, (7) airport and (8) railway. For the training set, only noise types (1)~(4) were added. For the test set, all 8 noises were added and further divided into three test sets, i.e., Set A, B and C. Test set A includes noise types (1)~(4), and set B has noise types (5)~(8). On the other hand, test set C has the noise types (1) and (5) but with different communication channels [13]. For each type of noise, 5 sets of noisy samples with signal-to-noise rations (SNRs) ranging from 0~20 dB were used for evaluation.

It is noted that set A, B, and C represent the scenario of seen and unseen noisy environment, and unseen communication channels, respectively.

Table 1. List of non-speech sources in RWCP Sound Scene Database in Real Acoustical Environments.

	Category	# of types	# of samples
Collision	Wood	12	1187
	Metal	10	1000
	Plastic	6	550
	Ceramic	8	800
Action	Article dropping	2	200
	Gas jetting	2	200
	Rubbing	5	500
	Bursting and breaking	2	200
	Clapping sound	10	829
Characteristic	Small metal articles	15	1072
	Paper	4	400
	Musical instrument	11	1079
	Electronic sound	8	705
	Mechanical	10	1000

4.3. Performance evaluation and comparison

In all the following experiments, 23-dimensional mel-filterbank features with 3 ms. frame shift were extracted. Then input super-vectors were made up of 10 neighboring frames, i.e., 230 dimensions to represent the spectro-temporal information of audio signals. Beside, HMMs with 7 states and 4 mixtures were adopted as the audio recognizers.

Firstly, two baseline systems were built using conventional 39-dimensional MFCCs features without and with MVA feature normalization (MFCC and MFCC+MVA). Their performances, 5.55% and 4.31% average recognition error rates (RERs) over test set A, B and C are shown in Fig. 4. These results indicate that feature normalization is very important for MFCCs to alleviate the interference of background noises.

Secondary, a large set of 350 Gabor filters were chosen by knowledge-based approaches [5]. The extracted spectro-temporal features were further analyzed and empirically reduced into 150-dimensional features using PCA. Fig. 4 shows that Gabor filter features (Gabor+PCA) achieved 4.43% RER. It is found that Gabor+PCA features work much better than MFCCs in noisy environment.

Thirdly, LDA method was also investigated using the same experimental settings as in the Gabor+PCA case. From Fig. 4, it is found that the performance of LDA features (4.20%) is slightly better than MFCC+MVA features.

Finally, the proposed MCE-based approach (initialized by PCA) was evaluated using also the same experimental settings. The results in Fig. 4 show the proposed new features are even more robust than LDA features and achieved the lowest RER of 3.17%.

Moreover, Fig. 5 shows the average recognition rates of the 5 features over (a) different SNRs and (b) test sets (Set A. Set B and Set C). From Fig. 5, it can be found that the proposed MCE-based features are superior to others in all cases. Besides, although LDA is the second best feature, it didn't work very well at high SNR (20 dB and 15 dB) and Set C. It may indicate

that LDA features are more sensitive to SNR and, especially, channel mismatch. Fig. 5 (a) also shows that Cabor+PCA features are better than MFCC+MVA in high SNR conditions but it degrades more quickly in low SNR cases.

Therefore, the results in Fig. 4 and 5 confirm the superiority and robustness of the proposed MCE-based spectro-temporal feature extraction approach in noisy environment.

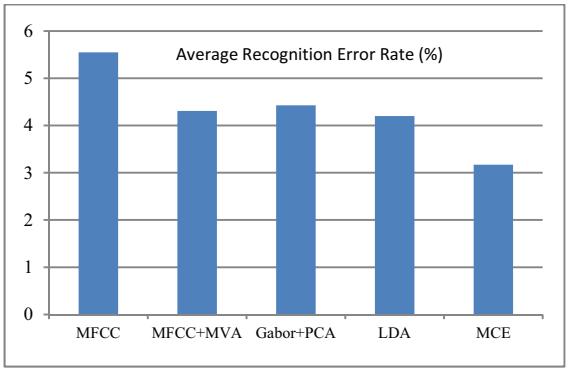

Fig. 4: Performance comparison of 5 different features on the noisy RWCP non-speech database.

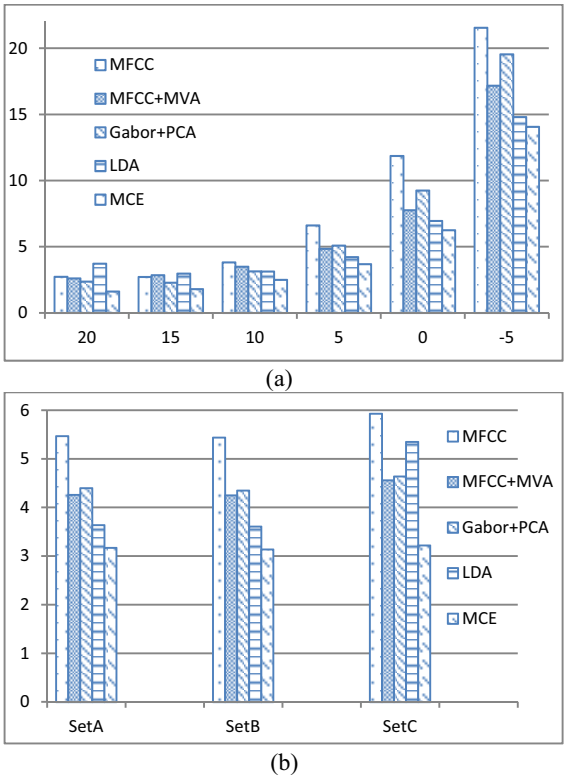

Fig. 5: Average recognition error rates (%) of 5 features over different (a) SNRs and (b) test sets on the noisy RWCP non-speech database.

4.4. Analysis

Fig. 6 shows the first 10 spectro-temporal filters found by the proposed MCE-based approach. It can be seen from Fig. 6 that some filters are similar to conventional Gabor filters or edge detectors. But others (the last three ones in the right-bottom panels) indeed have more complex spectro-temporal patterns. This may explain why the proposed new features are superior to Gabor+PCA features. In other words, the proposed method could capture more complex spectro-temporal cues.

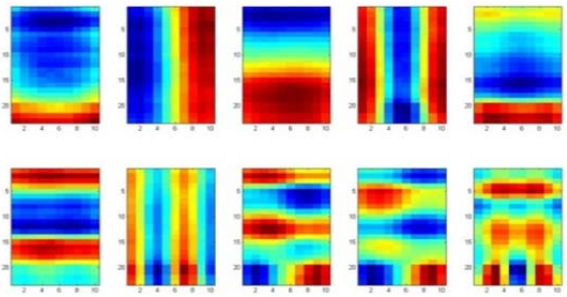

Fig. 6: The first 10 spectro-temporal filters (23 bands*10 frames) found by the proposed MCE-based approach.

5. Conclusion

In this paper, a MCE-based approach has been proposed to extract new spectro-temporal features as alternatives to MFCCs in order to meet the goal of audio classification. Experimental results on noisy RWCP non-speech database show that the proposed MCE-based features are superior to several state-of-the-art features. Besides, experiment analysis shows the proposed method could capture more complex spectro-temporal cues. We thus confirm the robustness of the proposed MCE-based spectro-temporal feature extraction method.

6. Acknowledgment

This paper is partially supported by National Science Council, Taiwan with project 97-2628-E-027-003-MY3 and 98-2221-E-027-081-MY3.

7. References

[1] Chia-Ping Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257-270, January 2007.

[2] Yuan-Fu Liao, Hung-Hsiang, Fang and Chi-Hui Hsu, "Eigen-MLLR environment/speaker compensation for robust speech recognition," in Proc. *InterSpeech*, 2008.

[3] Hermansky, H., Ellis, D. and Sharma, S., "Tandem connectionist feature extraction for conventional HMM system," in Proc. *ICASSP*, 2000.

[4] Depireux, D.A., Simon, J.Z., Klein, D.J. and Shamma, S.A., "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiology*, vol. 85, pp. 1220–1234, 2001.

[5] Zhao, S., Ravuri, S. and Morgan, N., "Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR," in Proc. *ICASSP*, 2009.

[6] Michael Kleinschmidt, "Robust speech recognition based on spectro-temporal processing," *PhD thesis*, 2002, Universitaet Oldenburg.

[7] K. Schutte and J. Glass, "Speech Recognition with Localized Time-Frequency Pattern Detectors," in Proc. *ASRU*, Kyoto, Japan, December 2007.

[8] Mesgarani, N.,Slaney, M. and Shamma, S., "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech and Language Processing*, Volume 14, Issue 6, p.920–930., 2006.

[9] Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Letter, 2010

[10] B. H. Juang, W. Chou, and C. H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*. Volume 5, No. 3, May 1997.

[11] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical Sound Scene Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," in Proc. *Int. Conf. Lang. Resources Evaluation*, 2000.

[12] H.G. Hirsch, D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in Proc. *6th International Conference on Spoken Language Processing - ICSLP'00*, 2000.

[13] H.G. Hirsch, "FaNT - Filtering and Noise Adding Tool," http://dnt.kr.hs-niederrhein.de/download/fant.tar.gz , 2010.