KERNEL WEIGHTED FISHER SPARSE ANALYSIS ON MULTIPLE MAPS FOR AUDIO EVENT RECOGNITION

Yu-Hao Chin[1], *Bo-Wei Chen*[2], *and Jia-Ching Wang*[1]

[1]Dept. of Computer Science & Information Engineering, National Central University, Taoyuan, Taiwan
[2]School of Information Technology, Monash University, Malaysia

ABSTRACT

This work presents a novel approach for audio event recognition. The approach develops a weighted kernel fisher sparse analysis method based on multiple maps. The proposed method consists of maps extraction and kernel weighted Fisher sparse analysis.

Two maps are firstly extracted from each audio file, i.e. scale-frequency map and damping-frequency map. The scale and frequency of the Gabor atoms are extracted to construct a scale-frequency map. On the other hand, the damping-frequency map is generated according to the frequency and damping factor of damped atoms. Gabor atoms can be utilized to model human auditory perception, and the damped atoms can be used to model commonly observed damped oscillations in natural signals. This work fuses the advantages of these two dictionaries to improve the performance of the system. During the recognition stage, this work constructs a kernel sparse representation-based classifier via the proposed kernel weighted Fisher sparse analysis to enhance separability. The proposed kernel weighted Fisher sparse analysis combines sparse representation with heteroscedastic kernel weighted discriminant analysis (HKWDA), which is useful for providing a discriminative recognition of audio events because a weighted pairwise Chernoff criterion is utilized in the kernel space. Experiments on a 20-class audio event database indicate that the proposed approach can achieve an accuracy rate of 82.70%. Also, integrating the scale-frequency map with MFCCs increases the accuracy rate to 87.70%.

Index Terms—Kernel weighted Fisher sparse analysis, scale-frequency map, damping-frequency map, kernel sparse classification, audio event classification

1. INTRODUCTION

Audio events refer to audio segments that present certain event scenarios or human-centered actions, which includes human speech and a wide range of non-speech sound classes, such as clapping, door knocking, and gunshot firing. Automatically audio events recognition has become an important issue because the technique can be applied to various applications [1]-[6].

Compared with the recognition of structured sound, such as music, audio event classification must deal with variant audio scenes and locations. Signal sources and surroundings may change frequently over time. Additionally, when desired signals are corrupted or even overwhelmed by inferences, sound data may lose their key features. Several studies [7, 8] have indicated that unlike structured signals, audios events may contain strong temporal features or broad flat spectra. Such a phenomenon could make conventionally adopted features, such as Mel-frequency cepstral coefficients (MFCCs) [9], linear predictive cepstral coefficients (LPCCs), and linear predictive coding (LPC) inapplicable to audio event classification [7] as these features were originally designed for modeling the spectral envelope of human vocal tracts [10, 11]. Besides, comparative analysis of acoustic models [12] revealed that both LPC and LPCCs linearly approximated sound over all frequencies. This is inconsistent with the perception of human hearing. LPC includes a large portion of the high frequency bands of a speech in which contains mostly noise. This inclusion of noise information may affect system performance. [13]. Although perceptual linear predictive (PLP) features [14] and MFCCs modified linear spectral distortion of LPCCs by employing psychophysically based Bark-scale and Mel-scale transformations respectively, the performance is still limited. Therefore, how to provide an algorithm that is feasible for processing unstructured signals is of priority concern.

Our previous work [15] discovered that the nonuniform scale-frequency map performs well in the task of sound recognition for home automation. However, it is not enough to analyze signal by only one dictionary because the acoustic properties of various audio events are different from each other. To solve the problems in the work of [15], this study presents a novel approach, which combines kernel weighted Fisher sparse analysis with scale-frequency maps and damping-frequency maps to classify audio events. The

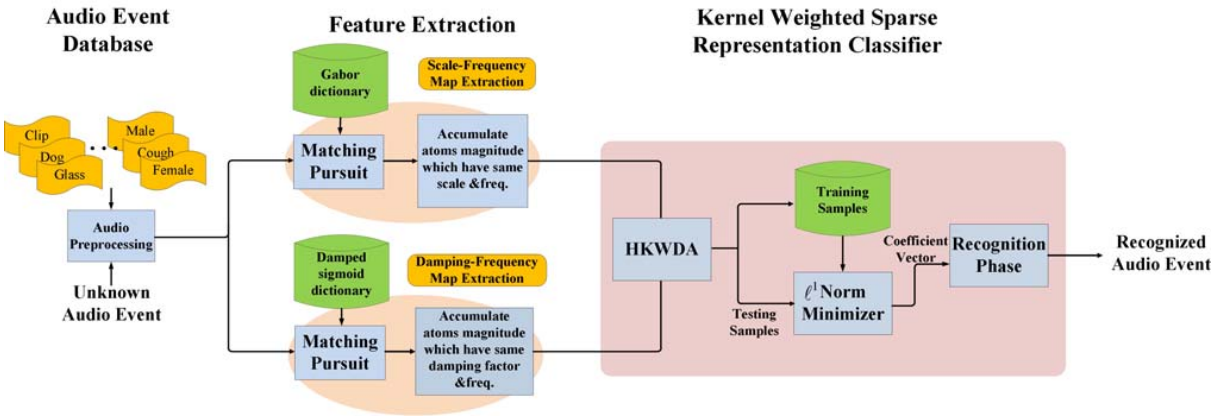

Fig. 1. Block diagrams of the audio events recognition system.

proposed method is characterized by converting multiple nonuniform maps into high-dimensional kernel space for sufficiently sparse coefficients, using the kernel weighted Fisher discriminant criterion. This method integrates both advantages of two maps. The scale-frequency maps and damping-frequency maps provide human auditory perception and damped oscillations analysis on critical bands, respectively, especially at low frequencies. Both of two maps can enhance separability during classification. Besides, this paper further utilizes kernel weighted Fisher sparse analysis to offer a discriminative recognition of audio events.

Figure 1 illustrates the system block diagram. Firstly, there are two maps extracted from each audio file, including scale-frequency map and damping-frequency map. With the use of a Gabor atom dictionary, the system can select several atoms to approximate the input signal by using the matching pursuit (MP) [19] method. Each atom in the dictionary takes the form of a Gabor function, which consists of frequency, scale, phase, and position information. Scale-frequency maps are subsequently generated according to the frequency and scale. In the work of damping-frequency map, the atoms are selected from the damped sigmoid dictionary instead of selecting from Gabor atom dictionary, and then the damping-frequency map is generated according to the frequency and damping factor. The damping-frequency map is subsequently combined with the Gabor dictionary-based scale-frequency map that mentioned before. Next, the two maps are mapped into high-dimensional space by a kernel function, forming bases of different classes.

During the classification stage, the kernel weighted Fisher sparse analysis transforms the two maps of an unknown input into combinational bases of one class by using Heteroscedastic Kernel Weighted discriminant Analysis (HKWDA) [16] and ℓ^1-norm minimization [17]. The class with the minimum combinational error is subsequently selected as the output.

The rest of this paper is organized as follows. Section 2 illustrates the construction of dictionaries. Section 3 summarizes the performance of the proposed method and the analysis results. Conclusions are finally drawn in Section 4.

2. DICTIONARIES CONSTRUCTION

2.1. Gabor Dictionary Based on Critical Frequency Bands

In this subsection, the detail of dictionary and the Gabor-based scale-frequency map are explained, which are briefly mentioned before. In MP algorithm, the choice of dictionaries significantly impacts the sound categorization ability. Chu *et al*. [7] have evaluated several dictionaries such as Fourier, Haar, and Gabor functions for their performance in sound classification. After many works, they have discovered that the Gabor dictionary could yield better results than the other dictionaries. This paper follows the suggestion of [15] to calculate scale-frequency maps.

Let ρ denote the scale, which controls the width of the Gabor function;
u represent the central temporal position of the Gabor function;
f refer to the frequency;
θ denote the phase;
t represent the time index;
K refer to the normalization factor such that $\|G_{\rho,u,f,\theta}\|^2 = 1$.

The Gabor function can be expressed as

$$G_{\rho,u,f,\theta}(t) = \frac{K_{\rho,u,f,\theta}}{\sqrt{\rho}} e^{-\pi(t-u)^2\rho^{-2}} \cos\left[2\pi f(t-u)+\theta\right] \quad (1)$$

Here, this paper makes the following options in the construction of the Gabor dictionary. The following parameters are selected to generate the Gabor atoms: $\rho = \{2^j \mid j = 1,2,\ldots,8\}$, u = {0, 64, 128, 192}, f = {150, 450, 840, 1370, 2150, 3400, 5800}, $\theta = 0$, and t = 0–255. Totally, there are 224 atoms (7 frequencies × 8 levels of scale × 4 central positions) in the dictionary.

Notably, the frequencies selected here for scanning atoms are based on the critical bands for human auditory perception [18, 21]. To restate, f is selected from the critical bands. Consequently, the Gabor dictionary is generated based on critical frequencies.

2.2. Damped Sinusoids Dictionary

This subsection describes the details of the damped sinusoids dictionary and damping-frequency map. Damped sinusoids attempt to model commonly occurring damped oscillations in natural signals [19]. Damped sinusoids are more appropriate than symmetrical Gabor atoms for representing transients. Therefore, the proposed method combines the damping-frequency map with the scale-frequency map. The damped sinusoids dictionary can take the following form

Let a denote the damping factor, which controls the width of the damped sinusoids function;
u represent the central temporal position of the damped sinusoids function;
ω refer to the frequency;
θ denote the phase;
n represent the time index;
S refer to the normalization factor such that $\|S_{\rho,u,f,\theta}\|^2 = 1$.

The damped sinusoids function can be expressed as

$$g^{+}_{\{a,\omega,\tau,\phi\}} = S_{\{a,\omega,\tau,\phi\}} a^{(n-\tau)} \cos[\omega(n-\tau)+\phi]u[n-\tau] \qquad (2)$$

Here, the following parameters are selected to generate the damped sinusoids: a = {0.11*j | j=2,3,…,9}, u = {0, 64, 128, 192},ω= {150, 450, 840, 1370, 2150, 3400, 5800}, $\theta = 0$, and t = 0–255. Totally, the dictionary contains 224 atoms (7 frequencies × 8 levels of scale × 4 central positions).

3. EXPERIMENTAL RESULTS

An audio database consisting of 20 classes was used for our experiments. Totally, there were 899 audio event clips, which cover various audio events, such as clapping, coughing, double clapping, female speeches, door knocking, laughing, male speeches, and screaming. The number inside the parentheses refers to the number of files in each class. Each clip lasted five seconds, and the sampling rate was 16 kHz with a resolution of 16 bits per sample. The frame size was 256 samples, with a 50% overlap in the two adjacent frames.

Each audio clip was divided into segments, with each one containing 16 frames. Finally, 50% of the dataset was used for training and 50% for testing by adopting cross-validation method.

3.1. Comparison of Different Dictionary Sizes in Terms of Classification Results

Before the other approaches are more thoroughly compared with each other, exactly which dictionary size is most effective for the proposed system is discussed first. Given that the value of a dictionary affects the solution of sparse coefficients, different dictionary sizes must be evaluated to obtain better classification results. The dictionary size evaluated in this subsection contains six values, i.e. 170, 227, 284, 341, 398, and 455. Each of these six values refers to a certain proportion of the entire training data set, i.e. 3/8, 4/8, 5/8, 6/8, 7/8, and 1. This work does not evaluate the two subsets refer to the 1/8 and 2/8 proportion of the entire training data set because the dictionary in sparse representation must be overcomplete. Besides, in this subsection, the proposed system adopts polynomial in HKWDA temporarily. According to Fig. 2, using the entire training data set to construct the dictionary is feasible for the proposed system.

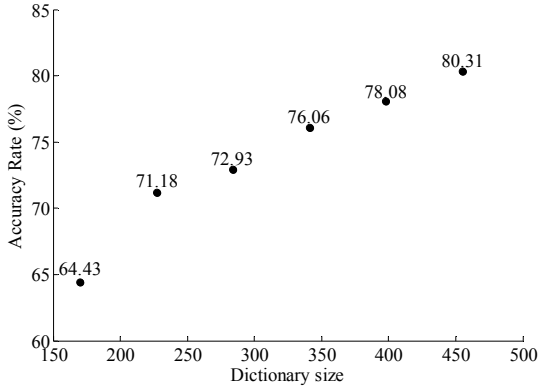

Fig.2. Comparison of different dictionary sizes in terms of accuracy rate.

3.2. Comparison of Various Kernels in HKWDA in Terms of Classification Results

This section evaluates the performance of the proposed system for various kernel types used in HKWDA. Four kernel types are tested: radial basis function kernel, polynomial kernel, linear kernel, and sigmoid kernel. Table 1 reveals that radial basis function kernel performs optimally for the proposed system.

Table 1. Comparison of different kernel types in terms of accuracy rate.

Kernel Type	Accuracy Rate (%)
Radial basis function	82.77
Polynomial	80.31
Linear	64.21
Sigmoid	64.42

3.3. Comparison of the Proposed Method and the Other Approaches

Recognition performance was assessed by designing an experiment to test the following approaches:

I. Frame-averaged MFCC (13) + SVM [20]: This baseline follows the idea of Temko's research [20] and uses a 13-dimensional frame-averaged MFCC feature. In addition, an RBF-kernel SVM is adopted in this system.

II. SFM+PCA+LDA+SVM [15]: This baseline follows the work of [15]. In this baseline, Principle Component Analysis (PCA) and LDA are applied to the scale-frequency map, subsequently generating the feature. During the classification phase, a segment-level multiclass SVM is operated.

III. Proposed system I - Two maps + sparse representation classification: This method uses the scale-frequency map and the damping-frequency map as features, and sparse representation classification is adopted.

IV. Proposed system II - Two maps + Frame-averaged MFCC + sparse representation classification: As same as the proposed system I. However, instead of proposed kernel sparse representation-based classifier, the traditional sparse representation classification is adopted.

V. Proposed system III - Two maps + kernel weighted fisher sparse analysis: As mentioned earlier, the proposed method consists of two processes. One is the extraction of two maps, and the other is kernel weighted fisher sparse analysis. When an unknown signal is input to the system, important atoms are extracted by using the matching pursuit algorithm. Scale-frequency map and damping-frequency map are firstly extracted from each audio file. Subsequently, the two mean maps based on all the frame-level maps are calculated. After feature extraction, the kernel weighted fisher sparse analysis based classification method is adopted.

VI. Proposed system IV – Two maps + Frame-averaged MFCC + kernel sparse representation-based classification: In this method, the scale-frequency map and damping-frequency map are concatenated with MFCCs. The kernel weighted fisher sparse analysis based classification method is also adopted in this system.

Table 2. Comparison between different methods.

Method	Accuracy Rate (%)
I: Frame-averaged MFCC + SVM [20]	74.50
II: SFM+PCA+LDA+SVM [15]	76.70
III: Proposed system I - Two maps + sparse representation classification	80.76
IV: Proposed system II - Two maps + Frame-averaged MFCC + sparse representation classification	83.67
V: Proposed system III – Two maps + kernel weighted fisher sparse analysis	82.77
VI: Proposed system IV - Two maps + Frame-averaged MFCC + kernel weighted fisher sparse analysis	87.70

Table 2 compares the results of the different approaches, where the first column of the table is the test approach, and the second is the accuracy rate. Clearly, the proposed system IV can achieve as high an accuracy rate as 87.70%. In comparison with methods I, II, III, IV, and V, the recognition rate of the method IV was increased by 13.20%, 11.00%, 6.94%, 4.03%, and 4.93%, respectively. Additionally, the proposed system IV also outperformed the other five methods by 8.02% on average.

4. CONCLUSION

This work has developed a novel audio event recognition approach. Firstly, the scale-frequency map and the damping frequency map are constructed and combined with each other. The scale-frequency map is constructed to model human auditory perception; the damping-frequency map is constructed to model the common occurrence of damped oscillations in natural signals. Combining these two maps provides more detailed information on signal characteristics at low frequencies. To further improve the performance of the system, this paper utilizes kernel weighted Fisher sparse analysis to enhance separability. Six analyses were performed to evaluate the effectiveness of the proposed method. The analysis results indicated that the overall accuracy could reach as high as 87.70%, i.e. significantly higher than the other five approaches.

In contrast with the baselines, experimental results demonstrated that the proposed method is more appropriate for audio event recognition. Our results further verified the performance of the proposed system, as well as the feasibility of the proposed algorithm.

5. REFERENCES

[1] R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for auditory scene categorization," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 596–606, Jun. 2008.

[2] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 713-719, Aug. 2011.

[3] P. K. Atrey and A. El Saddik, "Confidence evolution in multimedia systems," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1288–1298, Nov. 2008.

[4] J. Wang, E. Chng, C. S. Xu, H. Q. Lu, and Q. Tian, "Generation of personalized music sports video using multimodal cues," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 576–588, Apr. 2007.

[5] M. Cristiani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[6] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523-535, Oct. 2010.

[7] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[8] S. P. Ebenezer, A. Papandreou-Suppappola, and S. B. Suppappola, "Classification of acoustic emissions using modified matching pursuit," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 3, pp. 347–357, 2004.

[9] J. C. Wang, J. F. Wang, and Y. S. Weng, "Chip design of MFCC extraction for speech recognition," *J. VLSI Integration*, vol. 32, no. 1–2, pp. 111–131, Nov. 2002.

[10] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 181–184, Mar. 2007.

[11] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proc. Int. Conf. Acoustic, Speech and Signal Processing.*, Mar. 2008, pp.4733-4736.

[12] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 451–464, Sept. 1997.

[13] E. Wong and S. Sridharan, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification," in *Proc. Intell. Multimedia, Video Speech Process.*, 2001, pp. 95–98.

[14] H. Hcrmansky, N. Morgan, A. Bayya and P .Kohn, "Rasta-PLP Speech Analysis Technique," in *Proc. IEEE Int. Conf. international conference on Acoustics, speech and signal processing*, Apr. 1992, pp.121-124.

[15] J. C. Wang, C. H. Lin, B. W. Chen, and M. K. Tsai, "Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation, " *IEEE Trans. Automation Science and Engineering*, vol. 11, no. 2, pp. 607-613, Apr. 2014.

[16] G. Dai, D. Y. Yeung, and H. Chang, "Extending kernel Fisher discriminant analysis with the weighted pairwise Chernoff criterion," *9th European Conference on Computer Vision*, May. 2006, pp. 308-320.

[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[18] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[19] M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47, pp. 1890–1902, July 1999.

[20] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Proc. 1st Int. Evaluation Workshop on Classification of Events, Activities and Relationships*, Southampton, United Kingdom, 2006, Apr. 06–07, pp. 311–322.

[21] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.