소리 분류를 위한 NMF 특징 추출

조용춘^{O} 최승진 방승양
포항공과대학교 컴퓨터 공학과
{yongc^{O}, seungjin, sybang}@postech.ac.kr

NMF-Feature Extraction for Sound Classification

Yong-Choon Cho^{0} Seungjin Choi Sung-Yang Bang
Dept. of Computer Science, POSTECH

ABSTRACT

A holistic representation, such as sparse coding or independent component analysis (ICA), was successfully applied to explain early auditory processing and sound classification. In contrast, part-based representation is an alternative way of understanding object recognition in brain. In this paper, we employ the non-negative matrix factorization (NMF) [1] which learns parts-based representation for sound classification. Feature extraction methods from spectrogram using NMF are explained. Experimental results show that NMF-based features improve the performance of sound classification over ICA-based features.

1. INTRODUCTION

Classifying audio signals into speech, music, and environmental sounds is useful in audio retrieval system. Most of audio classification systems use frequency-based features or spectrum-based features. However because of its high dimensionality and significant variance for perceptually similar signals, direct spectrum-based features are not suitable. Recently Casey proposed an ICA-based sound recognition system which was adopted in MPEG-7 [2]. ICA is a statistical method which aims at decomposing multivariate data into a linear combination of non-orthogonal basis vectors with coefficient being statistical independent [3, 4]. Although ICA learns higher-order statistical structure of natural sounds (which leads to localized and oriented receptive field characteristics), it is a holistic representation because basis vectors are allowed to be combined with either positive or negative coefficients.

Parts-based representation is an alternative way of understanding the perception in the brain and certain computational theories rely on such representations. For example, Briederman claimed that any object can be described as *a configuration of perceptual alphabet* which is referred to as *geons* (geometric ions) [5]. An intuitive idea of learning parts-based representation is to force linear combinations of basis vectors to be non-subtractive. The NMF is a simple multiplicative updating algorithm for learning parts-based representation of sensory data.

In this paper, we propose methods of sound classification using NMF. Our sound classification systems extract non-negative component parts from spectro-temporal sounds as features. Basis vectors computed by NMF are re-ordered and portion of them are selected, depending on their discrimination capability. Sound features are computed from these reduced vectors and are fed into hidden Markov model (HMM) classifier. In addition, we also present a simple method of learning sound features which are robust to additive noise. We compare our methods with ICA-based method and confirm its high performance.

2. Non-Negative Matrix Factorization

The non-negative matrix factorization (NMF) is a subspace method which finds a linear data representation in non-negative constraint.

Suppose that N observed data points, $\{\mathbf{x}_t\}$, $t = 1, K, N$ are available. Denote the data matrix by $\mathbf{X} = [\mathbf{x}_1, L, \mathbf{x}_N]$. The latent variable matrix $\mathbf{S}$ is also defined in a similar manner. Under Poisson noise model, the log-likelihood is given by

$$L = \sum_{t=1}^{N}\sum_{i=1}^{m}\{\mathbf{X}_{it}\log(\mathbf{AS})_{it} - (\mathbf{AS})_{it}\} \qquad (1)$$

A local maximum of (1) is found by the following multiplicative updating rule (see [10] for details):

$$\mathbf{S}_{a\mu} \leftarrow \mathbf{S}_{a\mu} \frac{\sum_i \mathbf{A}_{ia}\mathbf{X}_{i\mu} /(\mathbf{AS})_{i\mu}}{\sum_k \mathbf{A}_{ka}}, \tag{2}$$

$$\mathbf{A}_{ia} \leftarrow \mathbf{A}_{ia} \frac{\sum_\mu \mathbf{S}_{a\mu}\mathbf{X}_{i\mu} /(\mathbf{AS})_{i\mu}}{\sum_\upsilon \mathbf{S}_{a\upsilon}} \tag{3}$$

The entries of $\mathbf{A}$ and $\mathbf{s}$ are all non-negative, and hence only non-subtractive combinations are allowed. This is believed to be compatible to the intuitive notion of combining parts to from a whole, and is how NMF learns a parts-based representation [1]. It is also consistent with the physiological fact that the firing rates are non-negative.

3. Feature Extraction by NMF

To extract the feature of an audio signal three steps are needed. Firstly, the audio signal is transformed to the spectrogram and this spectrogram is factorized by the NMF. Next, the basis matrix is ordered using the weight matrix, and finally, using the selected basis vectors, features for classification are extracted. Feature extraction is the procedure to factorize the new weight matrix from given ordered basis and spectrogram. Overall diagram is shown below.

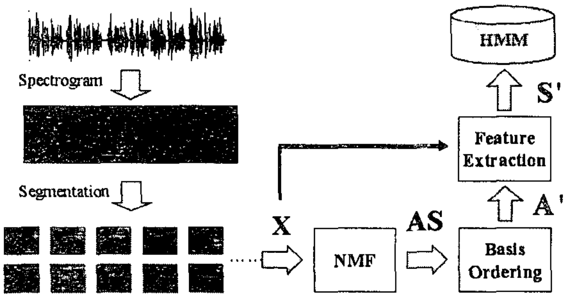

Fig. 1. Schematic diagram of our sound classification system.

3.1. NMF of an audio signal

To get a basis and weight matrix, we transform audio signal to the time-frequency domain via spectrogram. The spectrogram image is segmented by regular length. This segmentation procedure produces image patches. A image path is vectorized and NMF is directly applied to a set of vectors of image patches. In this procedure, an $m \times N$ matrix $\mathbf{X}$ (Spectrogram) is approximately factorized into an $m \times n$ matrix $\mathbf{A}$ and an $n \times N$ matrix $\mathbf{S}$. Usually n is chosen to be smaller than m or N, so that $\mathbf{A}$ and $\mathbf{S}$ are smaller than the original matrix $\mathbf{X}$. So it can be thought that matrix $\mathbf{A}$ contains the basis of data $\mathbf{X}$ and matrix $\mathbf{S}$ is the weight corresponding to $\mathbf{A}$.

3.2 Basis Ordering

For ordering the basis, we considered the discrimination power of each basis. It can be thought that the good feature has a good discrimination power. Therefore, we select the basis vectors that have an enough discrimination power. To do this, data distribution is considered to calculate the power of basis vectors.

$$J(k) = \sum_i \sum_j \frac{|m_{ik} - m_{jk}|}{\sigma_{ik} + \sigma_{jk}} \tag{4}$$

where, m_{ik} and σ_{ik} denotes the mean and variance of k^{th} row vector of matrix $\mathbf{S}$ that corresponds to class i. And $J(k)$ denotes the discrimination power of k^{th} basis. From this function, we can decide the appropriate threshold value to select some basis ($\mathbf{A}'$) which have enough discrimination capability.

3.3 Feature Extraction of Audio data

Although the NMF is linear, inference of the hidden representation $\mathbf{s}$ from a spectrogram $\mathbf{x}$ is highly non-linear because of its non-negative constraint. It is not clear how the best hidden representation could be computed directly from $\mathbf{A}$ and $\mathbf{x}$. However, as seen above, $\mathbf{s}$ can be computed by a simple iterative scheme. $\mathbf{A}$ can be regarded as constant, then only the update-equation for $\mathbf{S}$ remain (Method-I).

In this paper, weight matrix $\mathbb{S}$ was calculated by using the selected basis. Because the number of selected basis $\mathbb{A}'$ is always smaller than original basis $\mathbb{A}$, $\mathbb{A}$ matrix is updated partly ($\mathbf{A}''$) by (3) regarding $\mathbb{A}'$ as constant. Therefore we can take a new basis matrix $\mathbb{A}^{new}$ which is the concatenated matrix of $\mathbb{A}'$ and $\mathbb{A}''$. From this procedure, we can get a new $\mathbb{S}$ matrix, and the weights $\mathbf{S}'$, which is corresponded to $\mathbf{A}'$, is used for classification (Method-II).

$$\mathbf{A}''_{ia} \leftarrow \mathbf{A}''_{ia} \frac{\sum_\mu \mathbf{S}_{a\mu}\mathbf{X}_{i\mu} /(\mathbb{A}^{new}\mathbf{S})_{i\mu}}{\sum_\upsilon \mathbf{S}_{a\upsilon}} \tag{5}$$

$$\mathbf{A}^{new} = [\mathbf{A}', \mathbf{A}''], \qquad A'' \in ¡^{\,m\times(n-\kappa)}$$

$$\mathbf{S}_{a\mu} \leftarrow \mathbf{S}_{a\mu} \frac{\sum_i \mathbf{A}^{new}_{ia}\mathbf{X}_{i\mu} /(\mathbf{A}^{new}\mathbf{S})_{i\mu}}{\sum_k \mathbf{A}^{new}_{ka}} \tag{6}$$

$$\mathbf{S}' = [\mathbf{s}'_1, \mathbf{s}'_2, \mathrm{L}\ , \mathbf{s}'_N], \qquad \mathbf{s}' \in ¡^{\,\kappa}$$

This procedure makes $\mathbf{S}'$ noise robust because it allows the some new basis for noise and the sound which is not used in training.

5. Experiment

We used TIMIT database for speech, some commercial sounds for music and download sounds for musical instruments and environment sounds. The duration of the sound sequence was between 5 and 15 seconds. All sounds were resampled at 8KHz. Audio signals were transformed using short-time Fourier Transformation (STFT) computed for 25ms Hamming windows with a step size of 10ms for spectrogram. And we used a 100ms window with step size 50ms to produce an input signal for feature analysis of constant size. To do a NMF, the NMF equation (2) and (3) were iterated 200 times. We set 150 basis ($n=150$). We set 90% of threshold from cumulative function of (4), then we could get 113 ordered basis from 150 original basis. From this basis, we could extract feature (each column vector of $\mathbf{S}'$) by using (5) and (6).

Hidden Markov Model (HMM) classifier which had the 5-hidden states was used to test our system. And it trained a collection of 10 hidden Markov models using conventional maximum likelihood estimation (see (7) and Fig. 2)).

$$N^* \equiv \arg\{\max_{1\le j\le N} P(\mathbf{O}, I|\lambda_j)\} \qquad (7)$$

where, I denoted the most likely state sequence given observed data $\mathbf{O}$ and model parameters λ_j.

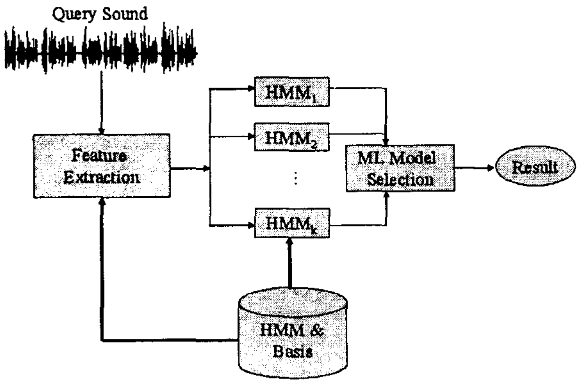

Fig. 2. Sound classification system

Table 1 shows the result for speech/music discrimination experiment of noisy data. All signals were added by 5 dB white noise.

Table 1. classification performance: Noisy data case

Class	Method-I		Method-II	
	correct	incorrect	correct	incorrect
Speech(Male)	30	0	30	0
Speech(Female)	13	17	25	5
Music	10	0	9	1
Total	53	17	64	6

Table 2 shows the comparison of two methods, NMF and ICA based method. ICA based classification was introduced in [2]. In our experiment, ICA based method was performed by conventional HMM for comparison. The performance for each method was measured as the percentage of correct classifications for the entire 126 test data. The result shows that the non-negative constraint is efficient to extract better features of audio data.

Table 2. Classification results for NMF and ICA

Class	NMF		ICA	
	# Hit	# Miss	# Hit	# Miss
Speech(Male)	30	0	30	0
Speech(Female)	30	0	28	2
Music	9	1	9	1
DogBark	9	0	2	7
Cello	10	0	9	1
Flute	9	1	9	1
Violin	7	0	2	5
Footsteps	9	0	8	1
Applause	3	2	2	3
Trumpet	4	2	5	1
Totals	120	6	104	22
Performance	95.24%		82.54%	

5. Conclusion

In this paper, we have shown that NMF based sound classificat on system and its performance. NMF based method yielded better performance than ICA based classification system using conventional HMM classifier. And it is shown that the sparse code is also effective in general sound classification.

REFERENCE

[1] D. D. Lee and H. S. Seung, " Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, Oct. 1999.

[2] M. Casey, " Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," in *Proc. Workshop on Consistent and Reliable Acoustic Cues for sound Analysis*, Eurospeech, Aalborg, Denmark, 2001.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., 2001.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, Inc., 2002.

[5] M. S. Gazzaniga, R. B. Ivry, and G. R. Mangum, *Cognitive Neuroscience: The Biology of the Mind*, W. W. Norton & Company, New York, 2001.