NON-NEGATIVE COMPONENT PARTS OF SOUND FOR CLASSIFICATION

Yong-Choon Cho, Seungjin Choi, Sung-Yang Bang

Department of Computer Science, POSTECH, Korea
{*yongc, seungjin, sybang*}*@postech.ac.kr*

ABSTRACT

Sparse coding or independent component analysis (ICA) which is a holistic representation, was successfully applied to elucidate early auditory processing and to the task of sound classification. In contrast, parts-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 in the task of sound classification. Methods of feature extraction from spectro-temporal sounds using the NMF in the absence or presence of noise, are explained. Experimental results show that NMF-based features improve the performance of sound classification over ICA-based features.

1. INTRODUCTION

Sound classification is an important problem in audio processing, which has many interesting applications. For example, speech/non-speech classification can be used to improve the performance of automatic speech recognition. Classifying audio signals into various types of sounds such as 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 direct spectrum-based features are not adequate in audio classification, because of its high dimensionality and significant variance for perceptually similar signals [2]. Recently Casey proposed an ICA-based sound recognition system which was adopted in MPEG-7 [2, 3].

ICA is a statistical method which aims at decomposing multivariate data into a linear combination of non-orthogonal basis vectors with coefficients being statistical independent [4, 5]. ICA was successfully applied to elucidate early auditory processing in the viewpoint of efficient encoding [6, 7] and was shown to well-match sparse auditory receptive fields [8]. ICA is a way of encoding sensory information efficiently and is a method of sparse coding, the usefulness of which was demonstrated in early visual processing [9] and in early auditory systems [10, 11]. 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) [12]. An intuitive idea of learning parts-based representation is to force linear combinations of basis vectors to be non-subtractive. The NMF [13] 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 the validity and high performance of our methods.

2. NON-NEGATIVE MATRIX FACTORIZATION

Efficient information representation plays a critical role in understanding perception of sensory data as well as in pattern classification. One way to elucidate an efficient coding strategy in early auditory processing is based on a linear generative model where the structure of the signals coming from external world is modelled in terms of a linear superposition of basis functions. In other words, the linear generative model assumes that the observed data $\boldsymbol{x}_t \in \mathbb{R}^m$ is generated by

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ contains basis vectors $\boldsymbol{a}_i \in \mathbb{R}^m$ in its columns, $\boldsymbol{s}_t \in \mathbb{R}^n$ is latent variable, and $\boldsymbol{\epsilon}_t \in \mathbb{R}^m$ is noise vector which represents uncertainty in the data model. Various methods for learning the linear generative model, include factor analysis, principal component analysis (PCA), sparse coding, and ICA. In general, these methods leads to holistic representation.

On the other hand, there is some evidence for parts-based representation in the brain, and certain computational theories of object recognition rely on such representations. One way to find parts-based representation using the linear generative model (1) is to constrain both basis vectors and latent variables to be non-negative so that non-subtractive combinations of basis vectors are used to model the observed data. The non-negative matrix factorization (NMF) [1] is a subspace method which finds a linear data representation in non-negativity constraint.

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

$$\mathcal{L} = \sum_{t=1}^{N} \sum_{i=1}^{m} \left\{ \boldsymbol{X}_{it} \log (\boldsymbol{AS})_{it} - (\boldsymbol{AS})_{it} \right\}. \tag{2}$$

A local maximum of (2) is found by the following multiplicative updating rule (see [13]) for details):

$$\boldsymbol{S}_{a\mu} \quad \leftarrow \quad \boldsymbol{S}_{a\mu}\frac{\sum_i \boldsymbol{A}_{ia}\boldsymbol{X}_{i\mu}/(\boldsymbol{AS})_{i\mu}}{\sum_k \boldsymbol{A}_{ka}}, \qquad (3)$$

$$\boldsymbol{A}_{ia} \quad \leftarrow \quad \boldsymbol{A}_{ia}\frac{\sum_\mu \boldsymbol{S}_{a\mu}\boldsymbol{X}_{i\mu}/(\boldsymbol{AS})_{i\mu}}{\sum_v \boldsymbol{S}_{av}}. \qquad (4)$$

The entries of $\boldsymbol{A}$ and $\boldsymbol{S}$ 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 from a whole, and is how NMF learns a parts–based representation [1]. It is also consistent with the physiological fact that the firing rate are non–negative.

3. FEATURE EXTRACTION BY NMF

Our methods of feature extraction from audio signals, consist of three steps. First, we compute spectrograms and segment them into a series of image patches through time to construct a data matrix $\boldsymbol{X}$ which is factored into a product of basis matrix $\boldsymbol{A}$ and the encoding matrix $\boldsymbol{S}$ by NMF. Next, a few number of basis vectors are selected, depending their discrimination capability. Finally features are learned using these selected basis vectors. The overall schematic diagram is shown in Fig. 1.

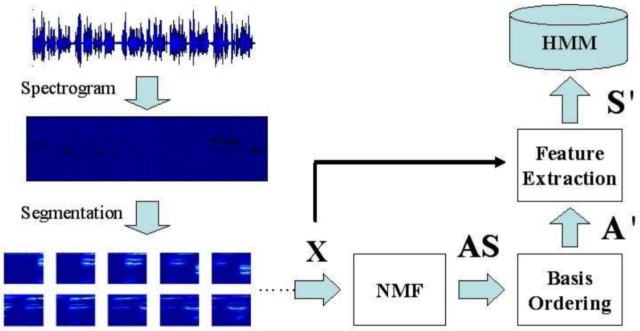
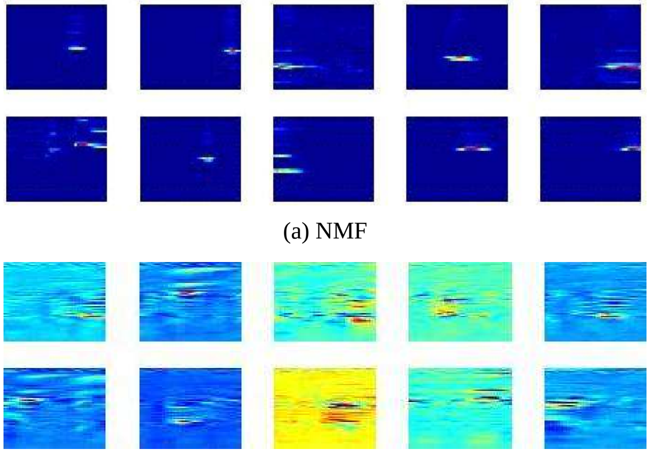


Fig. 1. Schematic diagram of our sound classification system.

3.1. NMF of Spectral Sounds

Audio signals sampled in time domain are transformed into spectrograms which represent time-dependent spectral energies of sounds. Spectrograms are segmented into a series of image patches through time. Hence, instead of working with time-domain audio signals, we play with a set of image sequence which do not allow negative values. Each image patch is converted into a vector and the data matrix $\boldsymbol{X}$ collects N vectors of dimension m. The NMF factors $\boldsymbol{X} \in \mathbb{R}^{m\times N}$ into a product of the basis matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and the encoding matrix $\boldsymbol{S} \in \mathbb{R}^{n\times N}$. The number of encoding variables (basis coefficients), n, is chosen to be smaller than the dimension of observation data, m. In other words, each image patch in spectrograms is modelled in terms of linear superposition of localized basis images with encoding variables representing the contributions of associated basis images. Exemplary basis images computed by NMF and ICA are shown in Fig. 2. NMF basis images exhibit much more localized characteristics than ICA basis images. Both NMF and ICA are inherently related to sparse coding, however, parts-based representation by NMF leads to more localized and sparse characteristics for non-negative data.



(a) NMF

(b) ICA

Fig. 2. 10 basis images (out of 150) learned by NMF and ICA from spectro-temporal sounds, are shown. Basis images learned by NMF show well-localized characteristics, compared to ICA basis images.

3.2. Basis Selection

NMF is an unsupervised learning method so that basis images are learned regardless of class labels. However, class information is available in a training phase, so it is desirable to take this information into account. Our basis selection scheme is based on the discrimination measure that is defined as

$$J(k) = \sum_i \sum_j \frac{|m_{ik} - m_{jk}|}{\sigma_{ik} + \sigma_{jk}}, \qquad 1 \leq k \leq n, \qquad (5)$$

where m_{ik} and σ_{ik} denotes the mean and variance of k^{th} row vector of matrix $\boldsymbol{S}$ that corresponds to class i. The discrimination measure (5) is reminiscent of Fisher's Linear Discriminant (FLD) measure which favors more separated mean and smaller variance. Fig. 3 shows that the discrimination measure of 150 basis vectors, where x-axis denotes the number of basis vector and y-axis denotes the magnitude of discrimination measure. By choosing an appropriate threshold value, we select a few number of basis vectors which are expected to have better discrimination. The matrix $\boldsymbol{A}'$ consists of $\kappa \leq n$ basis vectors selected by their discrimination measure.

3.3. Learning Features

NMF basis images show the auditory receptive field characteristics which are localized in frequency domain as well as in time domain (see Fig. 2). Although the NMF employs a linear data model, the inference of the hidden variable $\boldsymbol{s}$, given a basis matrix $\boldsymbol{A}$ and observed data $\boldsymbol{x}$, is a nonlinear process because of the non-negativity constraint. Hence, inferring an optimal hidden variable, given both $\boldsymbol{A}$ and $\boldsymbol{x}$ is not a trivial problem.

Here we investigate two different methods of inferring the best hidden variables (which will be auditory features), given that $\boldsymbol{A}' \in \mathbb{R}^{m\times\kappa}$ whose column vectors consist of κ basis vectors selected using the discrimination measure (5) from n basis vectors computed by NMF.

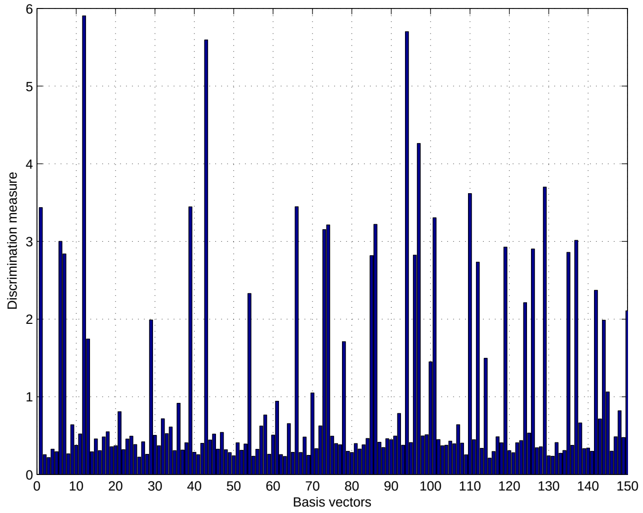


Fig. 3. The discrimination measure for each basis vector is shown.

Method I We compute the encoding variable matrix $\boldsymbol{S}'$ associated with $\boldsymbol{A}'$ by iterating the updating rule (3) until convergence with $\boldsymbol{A}'$ being fixed as constant.

Method II In Method I, only selected basis vectors were used to infer the associated encoding variables through the rule (3). In other words, $n - \kappa$ basis vectors (computed by NMF) do not make any contribution in inferring encoding variables. In contrast, Method II incorporate $n - \kappa$ basis vectors, $\boldsymbol{A}''$ into inferring the encoding variable matrix $\boldsymbol{S}'$. The basis matrix $\boldsymbol{A}$ is decomposed as $\boldsymbol{A} = [\boldsymbol{A}'\,, \boldsymbol{A}'']$ Only $\boldsymbol{A}''$ is updated while $\boldsymbol{A}'$ being fixed. Then partially updated matrix $\boldsymbol{A}$ is used to infer a new encoding variable matrix $\boldsymbol{S}$. Only the portion of $\boldsymbol{S} = [\boldsymbol{S}'^T \boldsymbol{S}''^T]^T$ associated with $\boldsymbol{A}'$, $\boldsymbol{S}'$, is kept for classification. This process is summarized below:

$$\boldsymbol{A}''_{ia} \leftarrow \boldsymbol{A}''_{ia} \frac{\sum_\mu \boldsymbol{S}_{a\mu} \boldsymbol{X}_{i\mu} / (\boldsymbol{A}^{new} \boldsymbol{S})_{i\mu}}{\sum_v \boldsymbol{S}_{av}} \tag{6}$$

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

$$\boldsymbol{S}_{a\mu} \leftarrow \boldsymbol{S}_{a\mu} \frac{\sum_i \boldsymbol{A}^{new}_{ia} \boldsymbol{X}_{i\mu} / (\boldsymbol{A}^{new} \boldsymbol{S})_{i\mu}}{\sum_k \boldsymbol{A}^{new}_{ka}} \tag{7}$$

$$\boldsymbol{S}' = [\boldsymbol{s}'_1\,, \boldsymbol{s}'_2\,, \cdots, \boldsymbol{s}'_N] \qquad \boldsymbol{s}' \in \mathbb{R}^\kappa$$

This procedure make feature $\boldsymbol{S}'$ noise robust, because it allows another basis for noise or other signals and weight $\boldsymbol{S}'$ is not corresponded to that basis. Table. 1 shows the noise robustness of this feature by the 3–classes test data which is included gaussian noise. In this table, classification performance is compared to each other. we could see that the Method-II was more noise robust.

4. HMM CLASSIFIERS

Hidden Markov models consist of three components; an initial state distribution π_i, a state transition matrix $\boldsymbol{T}_{ij}$ and the observation density function $b_j(\boldsymbol{o})$ for each state. Continuous HMMs set $b_j(\boldsymbol{o})$ to a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{K}_j$, giving $\boldsymbol{B}_j = \{\boldsymbol{\mu}_j, \boldsymbol{K}_j\}$ for each state. So, hidden Markov model j is denoted by $\lambda_j = \{\boldsymbol{T}_j, \boldsymbol{B}_j, \pi_j\}$ [14].

Table 1. Comparison of 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

For classification, a likelihood that measures the probability of each model given the observed data and the most likely state sequence $I = \{i_1, i_2, \ldots, i_T\}$ are estimated given observed data $\boldsymbol{O}$ and model parameters λ_j. The HMM classifier choose the model with the maximum likelihood score, among the N competing models.

$$N^* \equiv \arg\{\max_{1 \le j \le N} P(\boldsymbol{O}, I | \lambda_j)\} \tag{8}$$

5. EXPERIMENT

We used TIMIT database for speech, some commercial CDs for music and downloaded sounds for musical instruments and environment sounds. The duration of the sound sequence was between 5 and 15 seconds. The set of data was split into 40% training data and 60% test data.

In this experiment, all sounds were resampled at 8KHz. Spectrograms of sounds were computed using STFT with Hamming window of length 25 ms and overlapping of length 15 ms. Spectrograms were segmented through time using a window of length 100 ms shifted by 50 ms, in order to construct a data matrix. Using NMF updating rules (3) and (4), we computed 150 basis vectors ($n = 150$). These basis vectors were ordered, depending on its discrimination measure (5). We set a threshold in such a way that 90% of them were kept, hence, 113 ordered basis vectors were selected. We used these 113 basis vectors to infer encoding variables (features) using Method II.

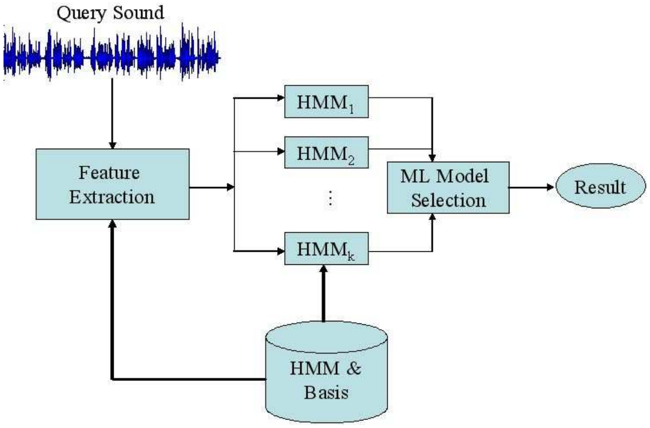


Fig. 4. Sound Classification System

For the test of our system, we used 10-class audio data and

HMM classifier that had the 5–hidden states and trained a collection of 10 hidden Markov models using conventional maximum likelihood estimation. To test the classifiers, the unseen data was presented to each HMM and the model with the highest likelihood was selected using (8) (see Fig. 4).

We performed two kinds of experiment, which were speech/music discrimination for noisy data and general sound classification experiment. For speech/music discrimination experiment for noisy data, HMM-classifier was trained by clean signal and performances were tested by noisy signal which was added by 5 dB white noise (see Method-II of Table. 1). For general sound classification experiment, we didn't consider the noise and the performance was compared to ICA based method using same training and test data.

Table. 2 shows the comparison of two methods, NMF and ICA method. ICA based classification was introduced in [3]. In our experiment, ICA based method was performed by conventional HMM for comparison. Correct classification were counted as *Hits*, and incorrect classifications were counted as *Missed*. The performance for each method was measured as the percentage of correct classifications for the entire 126 test data. The results show that both method has good performance; however, the NMF based method shows slightly better results than ICA based method. This result shows that the non-negative constraint is efficient to extract better feature of audio signals.

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%	

6. CONCLUSION

We have presented methods of feature extraction from spectral sounds, which are based on NMF. Compared to the ICA-based method, basis vectors computed by NMF showed more localized characteristics, which are close to parts-based representation. Methods of inferring encoding variables (corresponding features), given basis vectors and data, were proposed. In addition to localized characteristics of NMF basis vectors, our proposed basis selection scheme improved the classification performance. Using a conventional HMM classifier, we confirmed that our proposed methods outperformed the method based on ICA.

7. ACKNOWLEDGMENT

YC would like to thank Heeyoul Choi and Sangki Kim for their useful discussion. This work was supported by Korea Ministry of Science and Technology under Brain Science and Engineering Research Program and by BK 21 in POSTECH.

8. REFERENCES

[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, "Sound classification and similarity tools," in *Introduction to MPEG-7: Multimedia Content Description Language*, B. S. Manjunath, P. Salembier, and T. Sikora, Eds. John Wiley & Sons, Inc., 2001.

[3] 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.

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

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

[6] A. Bell and T. Sejnowski, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, vol. 7, pp. 261–266, 1996.

[7] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[8] K. P. Körding, P. König, and D. J. Klein, "Learning of sparse auditory receptive fields," in *Proc. IJCNN*, Honolulu, Hawaii, 2002.

[9] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

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

[11] S. Shamma, "On the role of space and time in auditory processing," *TRENDS in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

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

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, vol. 13.

[14] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Processing Magazine*, vol. 3, pp. 4–16, 1986.