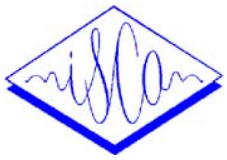
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Learning Nonnegative Features of Spectro-Temporal Sounds for Classification

Yong-Choon Cho

Digital Media R&D Center
Samsung, Korea
cho7573@hanmir.com

Seungjin Choi

Department of Computer Science
POSTECH, Korea
seungjin@postech.ac.kr

10.21437/Interspeech.2004-354

Abstract

In this paper we present a method of sound classification which exploits a parts-based representation of spectro-temporal sounds, employing the nonnegative matrix factorization (NMF) [1]. We illustrate a new way of learning nonnegative features using a variant of NMF and show its useful behavior in the task of general sound classification with comparison to independent component analysis (ICA) which produces holistic 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 recognizers. Classifying audio signals into various types of sounds such as speech, music, and environmental sounds, is useful in audio retrieval systems. A major challenge of general sound classification lies in selecting robust acoustic features and learning models with these features in such a way that diverse sound types are well classified. 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 encoding variables being statistical independent. ICA was successfully applied to elucidate early auditory processing in the viewpoint of efficient encoding [4] and was shown to well-match sparse auditory receptive fields [5]. 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. A 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) [6]. An intuitive idea of learning parts-based representations is to force linear combinations of basis vectors to be non-subtractive. The NMF [7] is a simple multiplicative updating algorithm for learning parts-based representations of sensory data.

In this paper we present a method of acoustic feature extraction from spectro-temporal sounds, which incorporates with a parts-based representation through the NMF. We first apply the NMF to the spectrogram of sounds in order to extract nonnegative basis vectors and associated encoding variables. These basis vectors are re-ordered and portion of them are selected, depending on their discrimination capability. Acoustic features are computed from these selected nonnegative basis vectors, are fed into a hidden Markov model (HMM) classifier. In addition, we also present a method of inferring encoding variables, given basis vectors learned by NMF. We compare our method with the ICA-based method and confirm the validity and high performance of our method.

2. Nonnegative Matrix Factorization

Suppose that N observed data points, $\{\boldsymbol{x}_t \in \mathbb{R}^m\}$, $t = 1, \ldots, N$ are available. Denote the data matrix by $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$. Linear model-based methods (such as PCA, ICA, and NMF) construct approximate factorization of the form

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{1}$$

where the columns of $\boldsymbol{A}\mathbb{R}^{m \times n}$ are basis vectors and each column of $\boldsymbol{S}\mathbb{R}^{n \times N}$ is called encoding variable vector (or latent variable vector).

Under the Poisson noise model, the log-likelihood is given by

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

where some irrelevant terms are left out. The NMF searches a local maximum of (2) by a multiplicative updating algo-

rithm which has the form

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

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

The entries of $\boldsymbol{A}$ and $\boldsymbol{S}$ are all nonnegative, 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. Instead of the maximum likelihood with the Poisson noise model, the I-divergence or the least squares criterion can be used in NMF, which leads to slightly different multiplicative updating algorithms [7].

3. Learning Features by NMF

Our methods of learning features from audio signals, consist of three steps. First, spectrograms of sounds are computed and are segmented into a series of image patches through time. Each image patch is converted to a vector through column-stacking, in order to construct a data matrix $\boldsymbol{X}$. Then the data matrix is factored into a product of a basis matrix $\boldsymbol{A}$ and an encoding variable 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 by our proposed inference scheme. The overall schematic diagram is shown in Fig. 1.

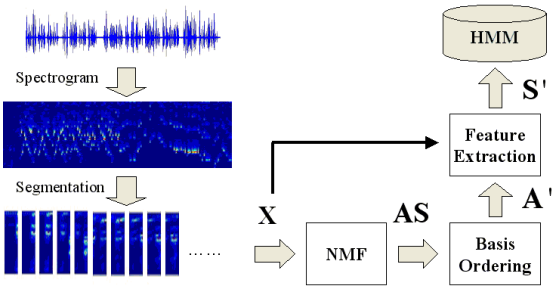
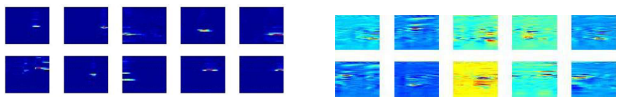


Figure 1: Schematic diagram of our sound classification system. The basis matrix $\boldsymbol{A}$ and the encoding variable matrix $\boldsymbol{S}$ are learned by NMF from the data matrix $\boldsymbol{X}$ which consists of the spectrogram patches in its columns. According to a discrimination measure, a few number of basis vectors are selected to construct a reduced-rank matrix $\boldsymbol{A}'$. Given $\boldsymbol{X}$ and $\boldsymbol{A}'$, a reduced-rank variable matrix $\boldsymbol{S}'$ (which corresponds to acoustic features), is computed by our proposed inference method. These features are fed into the HMM classifier.

3.1. Decomposing Spectro-Temporal Sounds by NMF

Audio signals sampled in the 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 by column-stacking, in order to construct a data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times N}$ which consists of N column vectors of dimension m where m corresponds the size of each image patch and N is the number of image patches. We decompose the data matrix $\boldsymbol{X}$ into a product of the basis matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and the encoding variable matrix $\boldsymbol{S} \in \mathbb{R}^{n \times N}$, using the NMF algorithm described in (3) and (4). 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, a parts-based representation by NMF leads to more localized and sparse characteristics for nonnegative data.



(a) NMF (b) ICA

Figure 2: Exemplary basis images learned by (a) NMF and (b) ICA from spectro-temporal sounds, are shown. Basis images learned by NMF show well-localized characteristics, compared to ICA basis images.

The basis matrix $\boldsymbol{A}$ and the encoding variable matrix $\boldsymbol{S}$ can be viewed as hair cells' receptive fields and vibrations, respectively. The encoding variables represent the degree of activation (responses) of hair cells, given a sound in the ear. Hair cells at the thick end, or the base of the cochlea are activated by high-frequency sounds and cells at the opposite ends, or the apex of the cochlea are activated by low-frequency sounds. These receptive fields have extensive overlap. So natural sounds like music or speech are made up of complex frequencies, thus, sounds activate a broad range of hair cells [6]. Therefore, it is desirable to select a set of appropriate bases (hair cells), for the sound classification task.

3.2. Basis Selection

NMF is an unsupervised learning method such that basis images are learned regardless of class labels. However, the 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 de-

fined by

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

where m_{ik} and σ_{ik} represent the mean and variance of the kth row vector of the matrix $\boldsymbol{S}$, in regards to the class i. The discrimination measure (5) is reminiscent of Fisher's Linear Discriminant (FLD) measure which favors more separated mean and smaller within-class variance. By choosing an appropriate threshold value, we select $\kappa \le n$ basis vectors which are expected to have better discrimination. A reduced-rank basis matrix $\boldsymbol{A}' \in \mathbb{R}^{m \times \kappa}$ is constructed by the κ basis vectors selected through their discrimination measure.

3.3. Learning Features: Inference of Encoding Variables

The basis images computed by NMF from the spectrograms of sounds, resemble the auditory receptive field characteristics, since they are well localized in the frequency domain as well as in the time domain (see Fig. 2). The basis selection method described in Sec. 3.2, produces a reduced-rank basis matrix $\boldsymbol{A}'$. Leaning acoustic features, given $\boldsymbol{A}'$ and $\boldsymbol{X}$, becomes a problem of finding associated encoding variables $\boldsymbol{S}'$. In PCA or ICA, encoding variables are easily computed by a linear mapping, i.e., $\boldsymbol{S}' = \left(\boldsymbol{A}'^T \boldsymbol{A}'\right)^{-1} \boldsymbol{A}'^T \boldsymbol{X}$. In contrast, the inference of encoding variables $\boldsymbol{S}'$ is a nonlinear process in NMF, due to nonnegativity constraints, although NMF is based on the linear data model. Therefore, it is not a trivial problem to infer optimal hidden variables (encoding variables), given $\boldsymbol{A}'$ and $\boldsymbol{X}$. Here we present two methods of inferring encoding variables (which correspond to sound 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. Two methods of inferring encoding variables are summarized below and a comparison between these two methods are shown in Figs. 3 and 4.

Method-I This is a simple way of inferring encoding variables, using the plain NMF updating rules with $\boldsymbol{A}'$ being fixed. In order to infer the encoding variable matrix $\boldsymbol{S}'$ associated with $\boldsymbol{A}'$, $\boldsymbol{S}'$ is updated until convergence using the rule (3), with $\boldsymbol{A}'$ being fixed.

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 $\boldsymbol{A} = \left[\boldsymbol{A}', \boldsymbol{A}''\right]$ where $\boldsymbol{A}' \in \mathbb{R}^{m \times \kappa}$ is the reduced-rank basis matrix (constructed from the basis selection) and $\boldsymbol{A}'' \in \mathbb{R}^{m \times (n-\kappa)}$ is a dummy matrix that takes part in inferring $\boldsymbol{S}'$. Only $\boldsymbol{A}''$ is updated while $\boldsymbol{A}'$ is fixed. Then partially updated matrix $\boldsymbol{A}$ is used to infer a new encoding variable matrix $\boldsymbol{S}$. Once this inference is done, only $\boldsymbol{S}'$ associated with $\boldsymbol{A}'$ where $\boldsymbol{S} = [\boldsymbol{S}'^T, \boldsymbol{S}''^T]^T$, is kept as features for classification. This inference process is summarized below:

$$\begin{aligned} \boldsymbol{A} &= [\boldsymbol{A}', \boldsymbol{A}''], \\ \boldsymbol{A}''_{ia} &\leftarrow \boldsymbol{A}''_{ia} \frac{\sum_\mu \boldsymbol{S}_{a\mu} \boldsymbol{X}_{i\mu} / (\boldsymbol{AS})_{i\mu}}{\sum_v \boldsymbol{S}_{av}}, \quad (6) \\ \boldsymbol{S}_{a\mu} &\leftarrow \boldsymbol{S}_{a\mu} \frac{\sum_i \boldsymbol{A}_{ia} \boldsymbol{X}_{i\mu} / (\boldsymbol{AS})_{i\mu}}{\sum_k \boldsymbol{A}_{ka}}. \quad (7) \end{aligned}$$

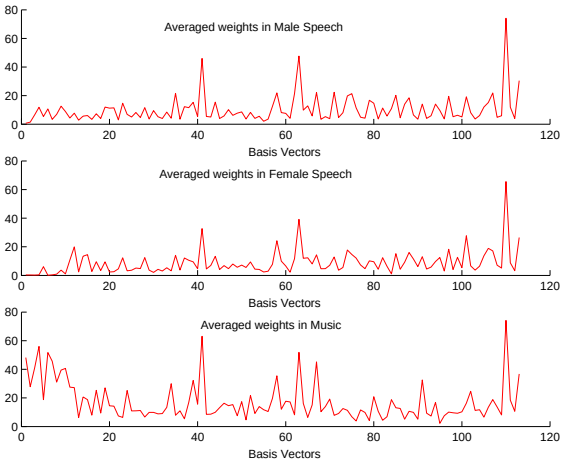


Figure 3: Distributions of averaged values of encoding variables associated with corresponding basis vectors in Method-I. Distributions for male speech, female speech, and music, are similar. For better discrimination, it is desirable for these distributions to be distinct.

4. Experiments

The sound data that we used in our classification experiments, consist of speech (from TIMIT), music (from some commercial CDs), musical instrument samples, and environmental sounds. The duration of sound signals was between 5 and 15 seconds. All sound signals were resampled at 8 kHz. The data was split into 40% for training sets and 60% for test sets.

Spectrograms 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. The NMF updating rules (3) and (4) were applied to compute 150 basis vectors ($n = 150$). These basis vectors were ordered, depending on their discrimination measure (5). For basis selection, we set up a threshold in such a way that 90% of basis vectors were kept, which produced 113 ordered basis vectors. We used these 113 basis

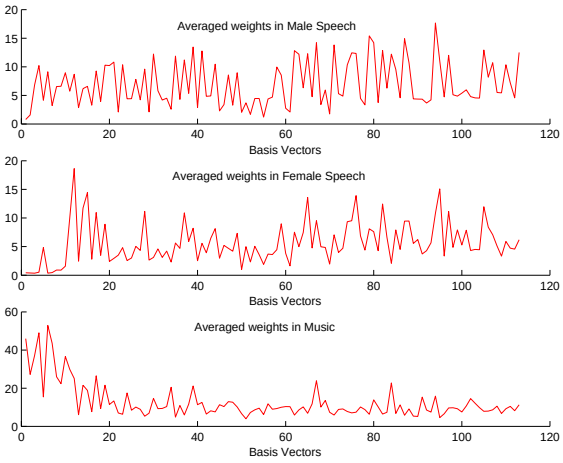


Figure 4: Distributions of averaged values of encoding variables associated with corresponding basis vectors in Method-II. Compared to Method-I, Method-II exhibits better discrimination for noisy data, since distributions show different characteristics.

vectors to infer encoding variables (features) using Method-I and Method-II.

We carried out a general sound classification experiment with 10 classes of audio signals. HMMs [8] were trained by a conventional maximum likelihood method and each HMM has 5 hidden states. In this experiment, we did not consider noisy data and compared our proposed method (Method-II) with an ICA-based method. Correct classification was counted by *Hits*, and incorrect classification was counted by *Missed*. The performance for each method was measured as the percentage of correct classification for the entire 126 test data. Table 1 summarizes the comparison results of our Method-II and the ICA-based method. Method-II outperforms the ICA-based method, which confirms the effectiveness of our new inference method and parts-based representations over holistic representations.

5. Conclusion

We have presented a method of sound classification which exploited parts-based representations of spectro-temporal sounds. The method used NMF to find nonnegative basis vectors, portion of which were chosen according to our discrimination measure. Given selected basis vectors, we have introduced a new way of learning nonnegative features. For classification, we have used a conventional HMM. Experimental results confirmed the high performance of our method, compared to ICA which provided holistic representations.

Table 1: Classification Results for Method-II and ICA

Class	Method-II		ICA	
	# Hit	# Miss	# Hit	# Miss
Speech (Male)	30	0	30	0
Speech (Female)	30	0	28	2
Music	9	1	9	1
DogBarks	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. Acknowledgment

This work was supported by Korea Ministry of Science and Technology under Brain Science and Engineering Research Program and by ETRI and BK 21 in POSTECH.

7. References

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

[2] M. Casey. Sound classification and similarity tools. In B. S. Manjunath, P. Salembier, and T. Sikora, editors, *Introduction to MPEG-7: Multimedia Content Description Language*. 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] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[5] K. P. Körding, P. König, and D. J. Klein. Learning of sparse auditory receptive fields. In *Proc. Int. Joint Conf. Neural Networks*, Honolulu, Hawaii, 2002.

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

[7] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

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