Nonnegative Features of Spetro-Temporal Sounds for Classification

Yong-Choon Cho, Seungjin Choi [*]

Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Abstract

A parts-based representation is a way of understanding object recognition in the brain. The nonnegative matrix factorization (NMF) is an algorithm which is able to learn a parts-based representation by allowing only non-subtractive combinations (Lee and Seung, 1999). In this paper we incorporate a parts-based representation of spectro-temporal sounds into the acoustic feature extraction, which leads to *nonnegative features*. We present a method of inferring encoding variables in the framework of NMF and show that the method produces robust acoustic features in the presence of noise in the task of general sound classification. Experimental results confirm that the proposed feature extraction method improves the classification performance, especially in the presence of noise, compared to independent component analysis (ICA) which produces holistic features.

Key words: Acoustic feature extraction, General sound recognition, Nonnegative matrix factorization

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

[*] Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
Email: seungjin@postech.ac.kr (S. Choi)

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 (Casey, 2001b).

Recently Casey proposed an ICA-based sound recognition system which was adopted in MPEG-7 (Casey, 2001b,a). 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 (Hyvärinen et al., 2001; Cichocki and Amari, 2002). ICA was successfully applied to elucidate early auditory processing in the viewpoint of efficient encoding (Bell and Sejnowski, 1996; Lewicki, 2002) and was shown to well-match sparse auditory receptive fields (Körding et al., 2002). 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 (Olshausen and Field, 1997) and in early auditory systems (Depireux et al., 2001; Shamma, 2001). 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) (Biederman, 1990). An intuitive idea of learning parts-based representations is to force linear combinations of basis vectors to be non-subtractive. The NMF (Lee and Seung, 2001) 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 show that the method produces robust acoustic features which improve the sound classification performance in the presence of noise. We compare our method with the ICA-based method and confirm the validity and high performance of our method.

2 Nonnegative 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 the 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 n basis vectors $\boldsymbol{a}_i \in \mathbb{R}^m$ in its columns, $\boldsymbol{s}_t \in \mathbb{R}^n$ is a latent variable vector, 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 representations.

On the other hand, there is some evidence for parts-based representations in the brain, and certain computational theories of object recognition rely on such representations. One way to find a parts-based representation using the linear generative model (1), is to constrain both basis vectors and latent variables to be nonnegative so that non-subtractive combinations of basis vectors are used to model the observed data. The nonnegative matrix factorization (NMF) (Lee and Seung, 1999) is a subspace method which finds a linear data representation with nonnegativity constraints.

Suppose that N observed data points, $\{\boldsymbol{x}_t\},\ t = 1, \ldots, N$ are available. Denote the data matrix by $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{m \times N}$. The latent variable (encoding variable, hidden variable) matrix $\boldsymbol{S} \in \mathbb{R}^{n \times N}$ is also defined in a similar manner. The NMF seeks to find a decomposition of the nonnegative data matrix, with nonnegativity constraints:

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ contains basis vectors in its columns and $\boldsymbol{S}$ is the encoding variable matrix with its each column being called an encoding representing the influence of each basis vector. Both $\boldsymbol{A}$ and $\boldsymbol{S}$ are restricted to have only nonnegative elements in the decomposition.

Various error measures for the factorization (2) with nonnegativity constraints, can be considered. Two exemplary error measures that were considered in (Lee and Seung, 2001) are summarized:

(1) (**LS**) The least square (LS) which employs the Euclidean distance between the data matrix $\boldsymbol{X}$ and the model $\boldsymbol{AS}$ leads to

$$\begin{aligned}\mathcal{E}_1 &= \|\boldsymbol{X} - \boldsymbol{AS}\|^2 \\ &= \sum_{i,j} \left[\boldsymbol{X}_{ij} - (\boldsymbol{AS})_{ij}\right]^2 .\end{aligned} \tag{3}$$

(2) (**I-divergence**) The I-divergence between $\boldsymbol{X}$ and $\boldsymbol{AS}$ leads to

$$\mathcal{E}_2 = \sum_{i,j} \left[\boldsymbol{X}_{ij} \log \frac{\boldsymbol{X}_{ij}}{(\boldsymbol{AS})_{ij}} - \boldsymbol{X}_{ij} + (\boldsymbol{AS})_{ij}\right] . \tag{4}$$

Note that I-divergence becomes identical to the Kullback-Leibler divergence when $\sum_{i,j} \boldsymbol{X}_{ij} = \sum_{i,j} (\boldsymbol{AS})_{ij} = 1$.

The minimization of the objective functions described above, should be done with nonnegativity constraints for both $\boldsymbol{A}$ and $\boldsymbol{S}$. Multiplicative updating is an efficient way in such a case, since it can easily preserve the nonnegativity for each iteration. Multiplicative updating algorithms for NMF associated with these two objective functions are given as follows:

(1) (**LS**) A local minimum of the objective function (3) is computed by the LS multiplicative algorithm that has the form

$$\boldsymbol{S}_{ij} \leftarrow \boldsymbol{S}_{ij} \frac{\left(\boldsymbol{A}^T \boldsymbol{X}\right)_{ij}}{\left(\boldsymbol{A}^T \boldsymbol{AS}\right)_{ij}}, \tag{5}$$

$$\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} \frac{\left(\boldsymbol{X}\boldsymbol{S}^T\right)}{\left(\boldsymbol{ASS}^T\right)_{ij}}. \tag{6}$$

(2) (**I-divergence**) For the case of I-divergence-based objective function (4), its minimum is found by the multiplicative updating algorithm that is of the form

$$\boldsymbol{S}_{ij} \leftarrow \boldsymbol{S}_{ij} \frac{\sum_k \left[\boldsymbol{A}_{ki} \boldsymbol{X}_{kj} / (\boldsymbol{AS})_{kl}\right]}{\sum_l \boldsymbol{A}_{li}}, \tag{7}$$

$$\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} \frac{\sum_k \left[\boldsymbol{S}_{jk} \boldsymbol{X}_{ik} / (\boldsymbol{AS})_{ik}\right]}{\sum_l \boldsymbol{S}_{jl}}. \tag{8}$$

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 More details on NMF can be found in (Lee and Seung, 1999, 2001).

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.

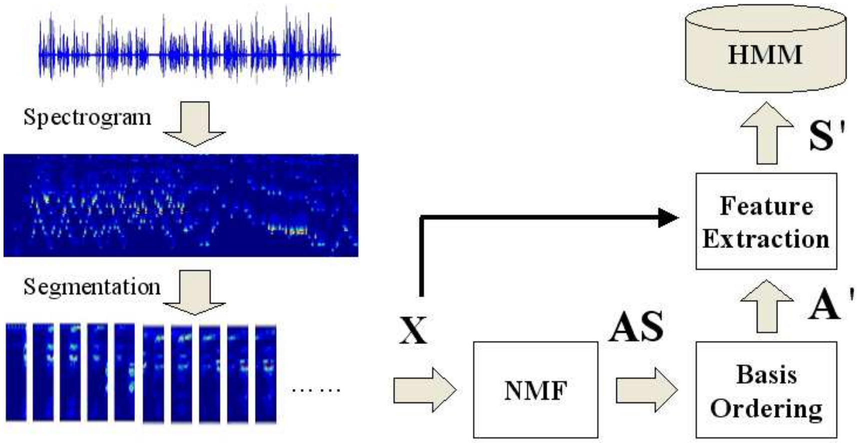


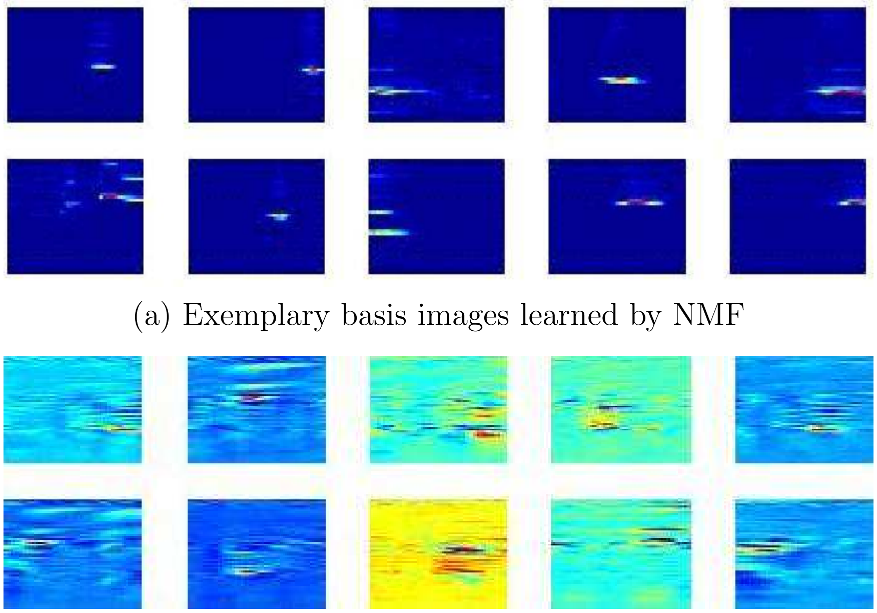
Fig. 1. Schematic diagram of our sound classification system. Spectrograms are segmented into a series of image patches through time. Each image patch is converted to a vector by column-stacking to construct a data matrix $\boldsymbol{X}$. For example, when the window size is 20 ms with 8 kHz sampling rate and 512-point-FFT is used, the size of image patch is 256×200. 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

In the view of neurophysiology, sound waves arriving at the ear make the eardrum to vibrate. These vibrations produce tiny waves within the inner ear's fluid that stimulate tiny *hair cells* located along the surface of the basilar membrane. Hair cells have receptive fields analogous to the retina's ganglion cells. While receptive fields of ganglion cells refer to a coding of locations in space, receptive fields of hair cells refer to a coding of sound frequency (Gazzaniga et al., 2001). Recently a sparse coding based on the skewness maximization (closely related to ICA) was successfully applied to elucidate the characteristics of sparse auditory receptive fields (Körding et al., 2002). In this paper we claim that a parts-based representation by NMF provides more localized basis vectors, compared to the sparse coding or ICA, in working with spectro-temporal sounds. Through the NMF method, very sparse localized auditory receptive fields can be computed.

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 (7) and (8). 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.

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 (Gazzaniga et al., 2001). Therefore, it is desirable to select a set of appropriate bases (hair cells), for the sound classification task.



(a) Exemplary basis images learned by NMF

(b) Exemplary basis images learned by ICA

Fig. 2. In our experiment, the size of each image patch is 256 × 200, leading to $m = 51200$. The reduced dimension n was chosen as 150. Exemplary 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 basis images computed by ICA.

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 defined by

$$\mathcal{J}(k) = \sum_i \sum_j \frac{|m_{ik} - m_{jk}|}{\sigma_{ik} + \sigma_{jk}}, \qquad 1 \le k \le n, \tag{9}$$

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 (9) is reminiscent of Fisher's Linear Discriminant (FLD) measure which favors more separated mean and smaller within-class variance. Fig. 3 shows the value of discrimination measure for 150 basis vectors computed by NMF, where the x-axis corresponds to basis vectors and the y-axis represents the value of the discrimination measure of associated basis vectors. Details on the experiment is described in Sec. 5. 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.

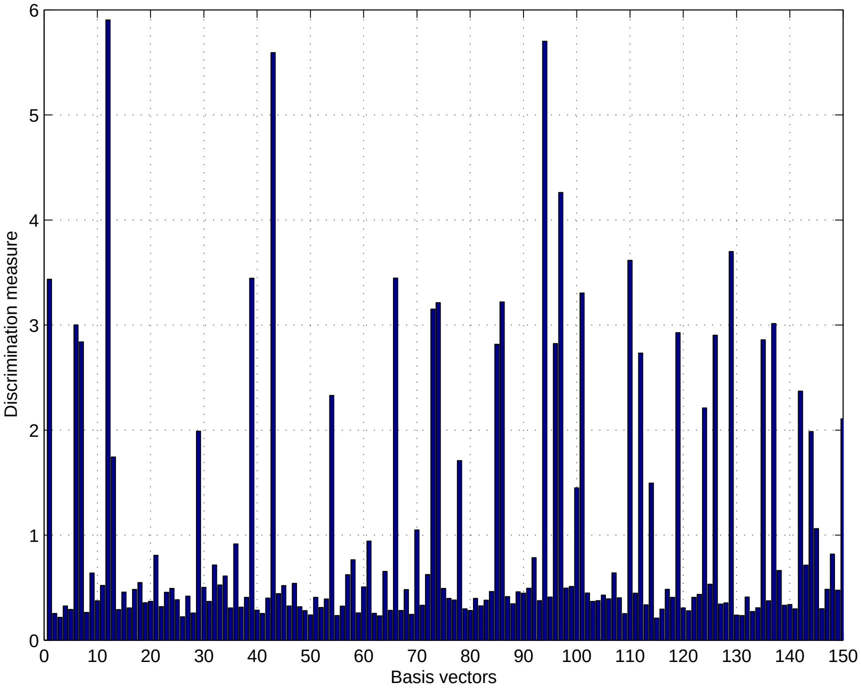


Fig. 3. For each basis vector, we compute the value of discrimination measure using the associated encoding variables, regarding the class labels. A few number of basis vectors show the high discrimination capability.

In contrast to a parts-based representation, this basis selection method in the task of sound classification, might be less critical for the case of holistic representations such as PCA and ICA, although it improves the classification performance slightly. Since holistic representations allow additive and subtractive combinations, fewer number of basis vectors could represent the data through a different combination without losing much information. On the other hand, parts-based representations come from non-subtractive combinations (especially in NMF). Thus the basis selection for a specific classification task plays an important role. Table. 1 summarizes a sound classification result, for the case of ICA-based features, with or without the basis selection method.

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 con-

Table 1
Comparison of classification performance for the case of ICA

Class	Without Basis Selection		With Basis Selection	
	# Hit	# Miss	# Hit	# Miss
Speech (Male)	30	0	30	0
Speech (Female)	29	1	30	0
Music	10	0	10	4
DogBarks	2	7	4	5
Cello	5	5	5	5
Flute	7	3	8	2
Violin	5	2	5	2
Footsteps	5	4	5	4
Totals	93	22	97	18

trast, 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 (9) from n basis vectors computed by NMF.

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 (7), 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 (7). 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 $\boldsymbol{A} = [\boldsymbol{A}', \boldsymbol{A}'']$ 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:

$$\begin{aligned}
&\boldsymbol{A} = [\boldsymbol{A}', \, \boldsymbol{A}''], \\
&\boldsymbol{A}''_{ia} \leftarrow \boldsymbol{A}''_{ia} \frac{\sum_{\mu} \boldsymbol{S}_{a\mu} \boldsymbol{X}_{i\mu} / (\boldsymbol{A}\boldsymbol{S})_{i\mu}}{\sum_{v} \boldsymbol{S}_{av}}, && (10) \\
&\boldsymbol{S}_{a\mu} \leftarrow \boldsymbol{S}_{a\mu} \frac{\sum_{i} \boldsymbol{A}_{ia} \boldsymbol{X}_{i\mu} / (\boldsymbol{A}\boldsymbol{S})_{i\mu}}{\sum_{k} \boldsymbol{A}_{ka}}. && (11)
\end{aligned}$$

Remarks

- Method-II produces more robust acoustic features in the presence of noise, compared to Method-I (see Table 2). Method-II allows a dummy matrix $\boldsymbol{A}''$ to be updated in inferring $\boldsymbol{S}$ with $\boldsymbol{A}'$ being fixed. Incorporating with a dummy matrix in inference, gives more flexibility to represent noisy data. In other words, noise or some unwanted signal components could be distributed in the encoding variables associated with the dummy matrix. After inference, $\boldsymbol{S}''$ is discarded and only $\boldsymbol{S}'$ is kept. Thus, Method-II produces more robust acoustic features.
- In fact, Method-II is a novel idea and is one of main contribution in this paper. Inferring encoding variables with a reduced number of basis vectors suffers from noisy data, since some irrelevant noise is also being captured by basis vectors. Hence, Method-II introduces dummy basis vectors (like an uncompression) in inferring encoding variables so that some irrelevant noise is captured by dummy basis vectors and discard these dummy basis vectors after inference.

4 HMM Classifiers

As a classifier, we use hidden Markov models (HMMs) that are trained by the acoustic features that are described in Sec. 3. Fig. 4 shows the structure of our sound classification system. Since HMM classifiers are well known, we briefly discuss them. See (Rabiner and Juang, 1986) for more details. HMMs consist of three components; an initial state distribution $\boldsymbol{\pi}$, a state transition matrix $\boldsymbol{T}$ and the observation density function $b(\boldsymbol{o})$ for each state. Continuous HMMs assume $b_j(\boldsymbol{o})$ to be a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{K}_j$. The jth HMM is described by $\lambda_j = \{\boldsymbol{T}_j, \boldsymbol{\mu}_j, \boldsymbol{K}_j, \boldsymbol{\pi}_j\}$. For the case of k classes, k HMMs are trained by their associated acoustic features. Given a test sound, its features are fed into k HMMs and likelihoods are computed. The HMM producing the highest likelihood is assigned as the class of the test sound.

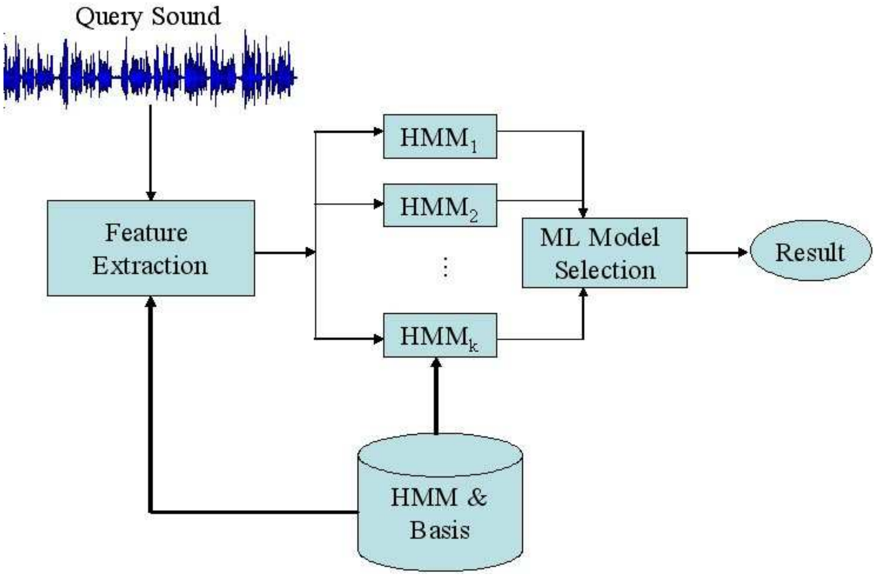


Fig. 4. The configuration of our sound classification system in shown. HMMs are trained by associated acoustic features. Each HMM correspond to a class. Given a query sound, associated features are computed and are fed into k different HMMs to computed the likelihood values. The class is assigned to the model that gives the highest likelihood value.

5 Experimental Results

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 (512 points) with Hamming window of length 25 ms and overlapping of length 15 ms. Spectrograms were segmented through time using a window of length 20 ms shifted by 10 ms, in order to construct a data matrix. The size of image patch is 256×200, leading to $m = 51200$. In order to choose the reduced dimension, n, we tried values between 100 and 200. For the dimension greater than 150, the overfitting phenomena was observed. Hence, from numerical study, we chose $n = 150$. The NMF updating rules (7) and (8) were applied to compute 150 basis vectors ($n = 150$). These basis vectors were ordered, depending on their discrimination measure (9). 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 vectors to infer encoding variables (features) using Method-I and Method-II.

Experiment 1 is involved with speech/music discrimination for noisy data. In this experiment, the HMM-classifier was trained by clean signals and was tested by noisy signals where 5 dB white noise was added to clean signals.

Averaged values of encoding variables (averaged activations) associated with corresponding basis vectors are shown in Figs. 5 and 6 for Method-I and Method-II, respectively. In Fig. 5, one can observe that Method-I produces similar distributions of activations of basis vectors for male speech, female speech, and music. On the other hand, Method-II gives some distinct distributions as shown in Fig. 6, which provides better discrimination. The classification result is summarized in Table 2).

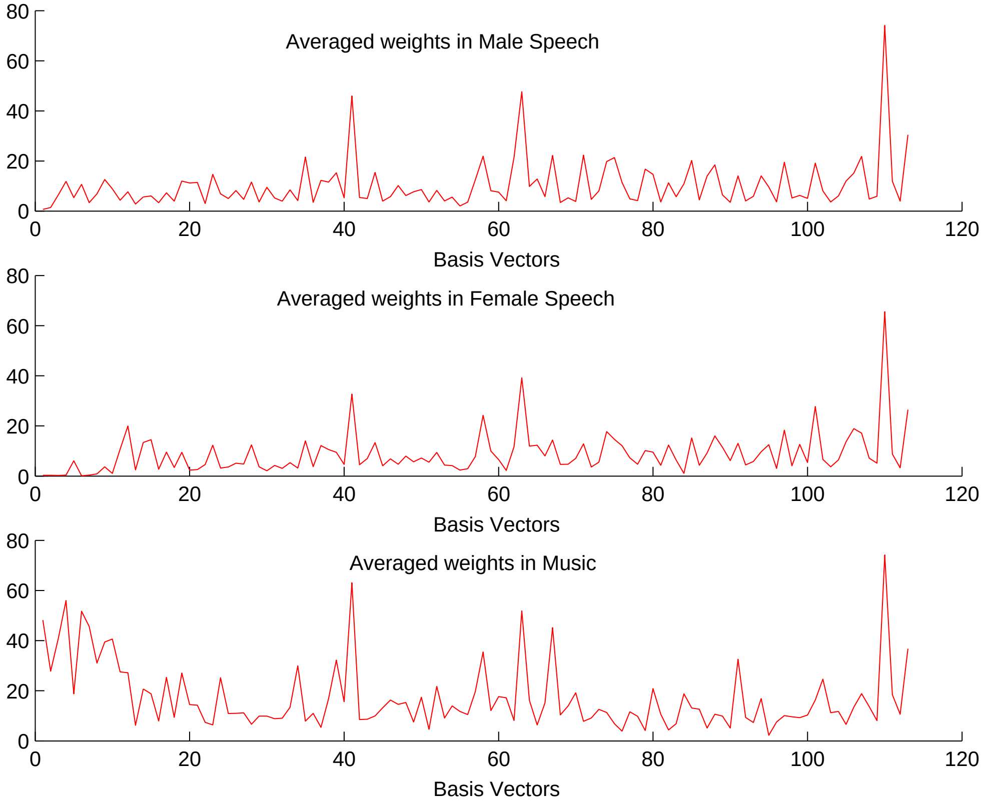


Fig. 5. Distributions of averaged values of encoding variables associated with corresponding basis vectors in Method-I. Distributions for male speech, female speech, and music, are similar. For better discrimination, it is desirable for these distributions to be distinct.

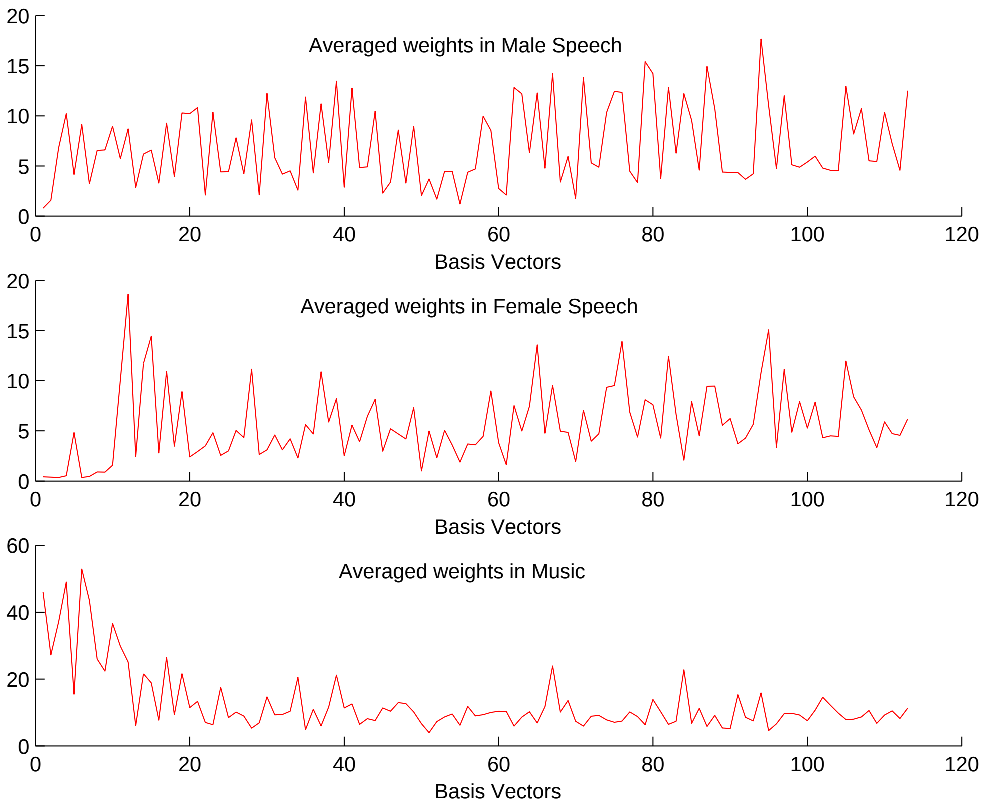


Fig. 6. Distributions of averaged values of encoding variables associated with corresponding basis vectors in Method-II. Compared to Method-I, Method-II exhibits better discrimination for noisy data, since distributions show different characteristics.

Table 2
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

5.2 Experiment 2

Experiment 2 is involved with the general sound classification. We carried out a sound classification experiment with 10 classes of audio signals. HMMs 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 3 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.

Table 3
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 Conclusion

We have employed NMF in order to extract nonnegative features from spetrotemporal sounds and have shown the useful behavior of these features, compared to ICA which gave holistic representations. Compared to the ICA-based method, basis vectors computed by NMF showed much more localized characteristics, which resembled sparse auditory receptive fields. We have presented a basis selection method and have introduced a new method of inferring encoding variables, given selected basis vectors in the framework of NMF. A key idea of this new inference method was to use dummy basis vectors which gave flexibility in representing noisy data. Acoustic features learned by our proposed inference method were shown to be robust, especially in the presence of noise. Together with the conventional HMM classifier, we confirmed that our proposed method outperformed the ICA-based method in the task of general sound classification.

7 Acknowledgment

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

References

Bell, A., Sejnowski, T., 1996. Learning the higher-order structure of a natural sound. Network: Computation in Neural Systems 7, 261–266.

Biederman, I., 1990. High-level vision. In: Osherson, D. N., Kosslyn, S. M., Hollberbach, J. M. (Eds.), Visual Cognition and Action: An Invitation to Cognitive Science. Vol. 2. MIT Press, Cambridge, MA.

Casey, M., 2001a. 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.

Casey, M., 2001b. Sound classification and similarity tools. In: Manjunath, B. S., Salembier, P., Sikora, T. (Eds.), Introduction to MPEG-7: Multimedia Content Description Language. John Wiley & Sons, Inc.

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

Depireux, D. A., Simon, J. Z., Klein, D. J., Shamma, S. A., 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiology 85, 1220–1234.

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

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons, Inc.

Körding, K. P., Körding, P., Klein, D. J., 2002. Learning of sparse auditory receptive fields. In: Proc. Int. Joint Conf. Neural Networks. Honolulu, Hawaii.

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

Lee, D. D., Seung, H. S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Vol. 13. MIT Press.

Lewicki, M. S., 2002. Efficient coding of natural sounds. Nature Neuroscience 5 (4), 356–363.

Olshausen, B. A., Field, D. J., 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1. Vision Research 37, 3311–3325.

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

Shamma, S., 2001. On the role of space and time in auditory processing. TRENDS in Cognitive Sciences 5 (8), 340–348.