SPEECH CODING AND NOISE REDUCTION USING ICA-BASED SPEECH FEATURES

Jong-Hwan Lee[1], *Ho-Young Jung*[2], *Te-Won Lee*[3], *Soo-Young Lee*[1]

[1]Brain Science Research Center and Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
373-1 Kusong-Dong, Yusong-Gu, Taejon, 305-701 Korea
(TEL: +82-42-869-8031, FAX: +82-42-869-8570, E-mail: jhlee@neuron.kaist.ac.kr)
[2]Electronics and Telecommunications Research Institute
161 Kajong-dong, Yusong-Gu, Taejon, 305-350, Korea
[3]Computational Neurobiology Laboratory, The Salk Institute
10010 N. Torrey Pines Road, La Jolla, California 92037, USA
and the Institute for Neural Computation, University of California, San Diego, USA

ABSTRACT

In this paper, we have obtained efficient speech features using independent component analysis to human speeches. When independent component analysis is applied to speech signals for efficient encoding the adapted basis vectors resemble Gabor-like features. Then only a few active coefficients of the trained basis vectors are sufficient for encoding the speech signals. Those trained speech features can be used in automatic speech recognition systems, and the proposed method gives better recognition rates than conventional mel-frequency cepstral coefficients (MFCCs) features. Trained basis vectors can be also applied for the removal of Gaussian noise. Speech signal corrupted by additive white Gaussian noise shows much improvements on the signal-to-noise ratio (SNR) after the denoising process. Then, these denoised speech features show better recognition performances than MFCCs features.

1. INTRODUCTION

Independent component analysis (ICA) was proposed as a method to solve the 'cocktail party problem'. ICA have taken a main interest in separating original independent source signals from the observed mixed signals. If natural images and natural sounds were composed of independent source signals, then the original source signals can be obtained from the observed image and sound signals using ICA. It means that ICA can be used for the feature extraction of image and sound signals. ICA had extracted feature vectors from natural scenes and music sound [1], [2]. In the feature extraction of natural images local and oriented edge filters could be obtained as a basis vectors [1]. These edge filters were very similar with the characteristics of V1 simple cell's receptive field on the visual cortex in our brain. And basis vectors trained from the natural sound were localized in both frequency and phase [2]. In relation to these works, speech features had been successfully extracted from human speech signals using ICA [3]. In that paper, the extraction of Gabor-like features from natural human speeches was reported. Extracted speech features look like bandpass filters which they have center frequencies and limited bandwidth. In this paper, to test the coding efficiency of ICA speech features simulation on reconstruction of speech signal using some selected speech features had been performed. And ICA speech features could be applied for automatic speech recognition systems. For each time frame, extracted feature coefficient vectors are obtained using trained basis vectors. Then, recognition rates with the ICA-based features are compared to those with the mel-frequency cepstral coefficients (MFCCs) for isolated-word recognition tasks. Trained basis vectors were also applied for the denoising of noisy speech signals. The coefficients of trained basis vectors have sparse distributions and maximum a posteriori (MAP) estimator could be used for denoising of the noisy speech signals corrupted by additive white Gaussian noise. Finally, using the denoised speech features noisy speech recognition experiments have been performed.

This research was supported as Brain Science & Engineering Research Program by Korean Ministry of Science and Technology.

2. EXTRACTING SPEECH FEATURES USING ICA

To extract independent feature vectors from speech signals, ICA algorithm is applied to a number of human speech segments. An ICA network is trained to obtain independent components $\mathbf{u}$ from the input speech segment $\mathbf{x}$. The trained weight matrix $\mathbf{W}$ extract basis vector coefficients $\mathbf{u}$ from $\mathbf{x}$. ICA assumes the observation $\mathbf{x}$ is the linear mixture of the independent source components $\mathbf{s}$. If $\mathbf{A}$ denote the inverse matrix of $\mathbf{W}$ then the columns of $\mathbf{A}$ represent basis feature vectors of observation $\mathbf{x}$.

$$\mathbf{u} = \mathbf{W} \cdot \mathbf{x}, \qquad \mathbf{x} = \mathbf{A} \cdot \mathbf{s} \tag{1}$$

To extract basis vectors one has to train mixing matrix $\mathbf{A}$ or unmixing matrix $\mathbf{W}$, and we trained the unmixing matrix $\mathbf{W}$. The learning algorithm is based on maximization of joint entropy $H(\mathbf{y})$

$$\Delta\mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \tag{2}$$

The outputs $\mathbf{y}$ are amplitude bounded random variables and therefore the marginal entropies are maximum for a uniform distribution of y_i. Finally, the information maximization algorithm is represented as [4]

$$\Delta\mathbf{W} \propto [\mathbf{W}^T]^{-1} + \left(\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})}\right)\mathbf{x}^T \tag{3}$$

where $p(\mathbf{u})$ denotes the approximation of the speech signal component probability density function, $p(u_i) = \partial y_i/\partial u_i = \partial g(u_i)/\partial u_i$. Here, $g(\mathbf{u})$ is a nonlinearity function, which approximates the cumulative distribution function of the independent source signal $\mathbf{u}$ [4].

Natural gradient is also introduced to improve a converging speed [5]. Particulary, this method does not require the inverse of matrix $\mathbf{W}$, and provides the following rule:

$$\Delta\mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W} = \left[\mathbf{I} - \varphi(\mathbf{u})\mathbf{u}^T\right]\mathbf{W}, \tag{4}$$

where $\varphi(\mathbf{u})$ is related to the source probability density function and called as the score function. If we assume that the speech signal components, $\mathbf{u}$ have a *Laplacian* distribution, $p(u) = exp(-|u|)$ then the score function becomes the $sign(\cdot)$ function. It improves the coding efficiency of speech signals, since most of the coefficients on $\mathbf{u}$ are almost zero and only a few important informations of speech signals are encoded in the tail of the *Laplacian* distribution.

Using the learning rule (4), $\mathbf{W}$ is iteratively updated by gradient ascent manner until convergence is achieved. Let's denote N as the size of speech segments, which are randomly generated from training speech signals. ICA network is composed of N inputs and N outputs, and N basis vectors are produced from $N \times N$ matrix $\mathbf{A}$ ($\mathbf{A} = \mathbf{W}^{-1}$).

3. FEATURE EXTRACTION USING REAL DATA

3.1. Selection of Dominant Feature Vectors

For the use of coding or recognition of speech signal, one may select dominant feature vectors from the N basis vectors. The ICA algorithm finds independent components corresponding to the dimensionality of the input, and may result in redundant components. To reduce this redundancy, several techniques have been proposed [6]. In this paper, the contribution of basis vectors to the speech signal and the variability of the basis vector coefficients were considered. We could see those two ordering methods provide almost same basis vector order [3]. Therefore, from N basis vectors ordered in decreasing variance of each basis vector coefficients, M dominant feature vectors can be selected.

3.2. Training Using Real Data and Recognition Experiments

To train the basis vectors from natural human speech signals, 75 phonetically balanced Korean words uttered by 59 speakers were used. Speech segments composed of 50 samples, i.e., 3.1ms time interval at 16kHz sampling rate, were randomly generated. Total 10^5 segments were generated, and each segment was pre-whitened to improve the convergence speed [1]. The whitening filter, $\mathbf{W}_z$ is:

$$\mathbf{W}_z = \langle(\mathbf{x} - \langle\mathbf{x}\rangle)(\mathbf{x} - \langle\mathbf{x}\rangle)^T\rangle^{-\frac{1}{2}} \tag{5}$$

This removes both first- and second-order statistics from the input data, $\mathbf{x}$ and makes the covariance matrix of $\mathbf{x}$ equal to $\mathbf{I}$.

Then, unmixing matrix $\mathbf{W}$ in ICA network was obtained by the learning rule (4) using those speech segments. $\mathbf{W}$ was initialized to an identity matrix, and $\varphi(\mathbf{u})$ was assumed as a sign$(\cdot)$ function for the efficient encoding. 300 sweeps through whole segments were performed, and $\mathbf{W}$ was updated every 100 input speech segments. The learning rate in (4) was fixed to 0.001 during the first 100 sweeps, 0.0005 during the next 100 sweeps, and 0.0001 during the last 100 sweeps.

Several learned basis vectors are localized in both frequency and time and resemble Gabor-like filters [3]. The average signal-to-noise ratio (SNR) between ICA basis vectors and fitted Gabor filters were 11.0dB.

To test the coding efficieny of the ICA basis vectors, simulation on reconstruction of the input speech signal had been performed. When 30 ICA basis vectors were selected by the variance order, SNR was 23.1dB and this result was superior than 20.6dB, that of the discrete Fourier basis functions. In the case of 20 ICA basis vectors, SNR was 15.6dB, and the recovered speech signals have a moderate quality of sound for listening. This result represents that the ICA basis vectors can be successfully applied for the efficient encoding of human speech signal than discrete Fourier basis functions.

The ICA-based features were applied to an isolated-word recognition task. The vocabulary consists of 75 Korean words, and 38 and 10 speakers uttered once to form training and test data, respectively. Whole word models were used for classification, and were represented by 15-state left-to-right continuous-density hidden Markov model (CDHMM). For comparison, standard MFCC features were extracted. In MFCC feature extraction process, filter bank which had 18 mel-scaled center frequencies were used. When 30 feature vectors were selected in the variance order, the proposed method yielded 36.8% error reduction compared to the case of the standard MFCCs. This result shows that the only a few active coefficients of $\mathbf{u}$ are sufficient for recognizing the speech signals [3].

4. FEATURE EXTRACTION IN NOISE ENVIRONMENT

4.1. Maximum a posteriori estimator

Trained basis vectors in section 3.2 were applied for the removal of Gaussian noise. In noise environment, if we denote y as a noisy coefficient of basis vector, s as a original clean coefficient of basis vector, and ν as a Gaussian noise with zero mean and σ^2 variance then

$$y = s + \nu \tag{6}$$

We want to estimate s from the only observed noisy coefficient y. Maximum a posteriori (MAP) estimation can be applied for this denoising process [9].

$$\hat{s} = \arg\max_s [P(s)P(y|s)] \tag{7}$$

If we denote p as a probability density function of sparse component s, then

$$\hat{s} = \arg\min_s [f(s) + \frac{1}{2\sigma^2}(y-s)^2] \tag{8}$$

where $f = -\log p$, the negative log-probability density function of s. Finally, MAP estimator gives this equation.

$$\hat{s} = h(y) \tag{9}$$

where the nonlinear function h is called as *shrinkage* function, and the inverse is given by

$$h^{-1}(s) = s + \sigma^2 f'(s) \tag{10}$$

Thus, the MAP estimator is obtained by inverting a certain function involving f', or the score function φ in (4) [9].

4.2. Denoising of noisy speech signal

The coefficients of the trained basis vectors in section 3.2 have sparse distributions since we assume that the each coefficients of basis vectors have *Laplacian* distribution. Then the trained ICA basis vectors can be applied to MAP estimator. To recover the denoised speech signal from the noisy speech signal several steps are needed. First, one needs to estimate shrinkage function h_i of the i-th basis vector. The same training data for the case of basis vector training can be used to estimate the shrinkage functions. Basis vectors and its shrinkage functions can be obtained at the same time. Second, calculate the noisy coefficient vector $\mathbf{y}$ for the noisy input speech signal $\tilde{\mathbf{x}}$, $\mathbf{y} = \mathbf{W}\tilde{\mathbf{x}}$. Third, obtain the denoised coefficients, $\hat{s}_i = h_i(y_i)$. Finally, recover the denoised speech signal, $\hat{\mathbf{x}} = \mathbf{W}^{-1}\hat{\mathbf{s}} = \mathbf{A}\hat{\mathbf{s}}$.

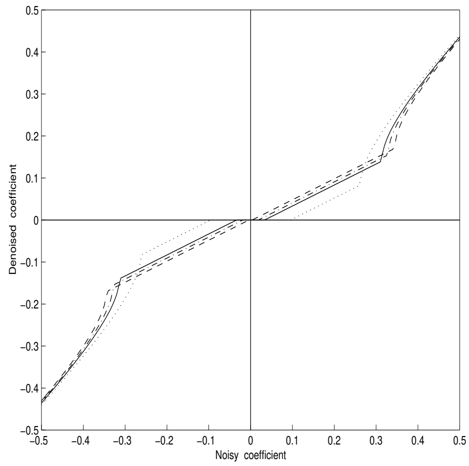
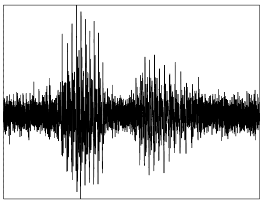


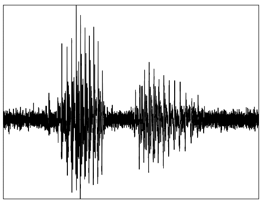
Figure 1: Shrinkage functions of four basis vectors. x axis is the noisy coefficient y_i, y axis is the denoised coefficient $\hat{s}_i$.

Fig.1 shows the estimated shrinkage functions of some basis vectors. Each shrinkage function has different shrinkage and scaling characteristics. For example, shrinkage function represented as dotted line makes the noisy coefficients zero through wider range than that of the solid line so the coefficients of the dotted line's basis vector has more sparse distribution than that of the solid line.

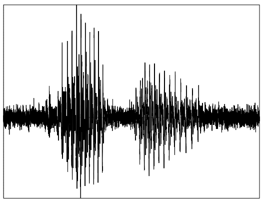
Noisy speech signals corrupted by additive white Gaussian noise were applied to denoising experiments. Speech data were the same as that of the basis vector training, 75 Korean words uttered by 59 speakers were used for the denoising. Fig.2 shows the waveforms of noisy speech signals corrupted by additive white Gaussian noise and denoised speech signals. We can see that the Gaussian noise in noisy speech is degraded to some extent. To measure the denoising capacity signal-to-noise ratio (SNR) was calculated before and after denoising.



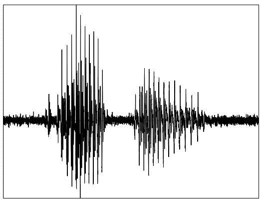
(a) Noisy speech (5dB AWGN)



(b) Denoised speech (5dB AWGN)



(c) Noisy speech (10dB AWGN)



(d) Denoised speech (10dB AWGN)

Figure 2: Noisy speech signal and denoised speech signal waveforms.

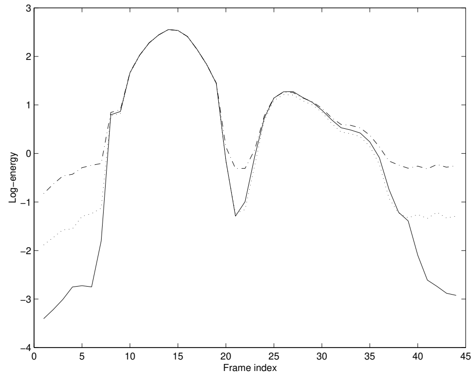
Table 1: Signal-to-noise ratio (SNR) results of the denoised speech signals.

	SNR (dB)			
Before denoising	5.0	10.0	15.0	20.0
After denoising				
$\epsilon = 0.001$	9.0	13.4	17.7	22.0
$\epsilon = 0.01$	9.8	13.9	17.9	22.2

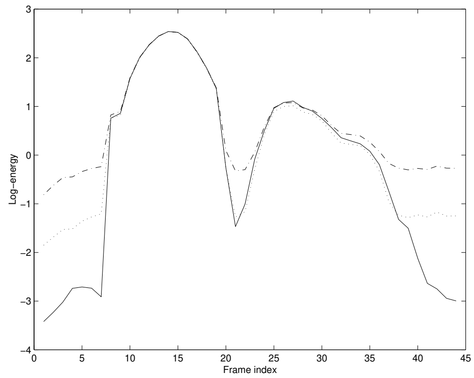
Table 1 shows the result of the denoising. In this table, the small constant value ϵ is required to estimate the shrinkage function. To get the value of the probability at zero of i-th basis vector coefficient s_i i.e. $p_{s_i}(0)$ one needs to assume the value of ϵ by small constant.

$$p_{s_i}(0) \approx E\{k(s_i)\}, \qquad k(s_i) = \begin{cases} 1/(2\epsilon) & \text{if } |s_i| < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

In the table 1 the results show that MAP denoising process with ICA basis vectors gives some improvements on SNR about 2∼4.8dB.



(a) A coefficients log-energy of the 6th basis vector



(b) A coefficients log-energy of the 7th basis vector

Figure 3: Log-energy spectrum transition of some basis vectors. Solid line: a log-energy of clean speech. Dash-dotted line: a log-energy of noisy speech signal with 15dB white gaussian noise. Dotted line: a log-energy of denoised speech signal.

4.3. Noisy speech recognition

Using the denoised basis vector coefficients $\hat{\mathbf{s}}$ in (9) noisy speech recognition experiments were performed. In the feature extraction of noisy speech signal, denoised basis vector coefficients can be obtained from the three steps in section 4.2. Then M log-energy spectrum can be obtained from the M denoised basis vector coefficients. Fig.3 shows the transition of the log-energy spectrum before and after denoising. Except for the non-speech parts in the beginning and the ending of speech signal, spectrum distortion could be decreased quite well.

Table 2 shows the recognition rates of noisy speech signals. Additive white Gaussian noises were mixed with clean speech signals and when SNR was 20dB, 15dB, 10dB, 5dB recognition rates were measured. Training and test speech data were the same as that of section3.2, and 15-state left-to-right CDHMM was also used as a classifier. Even though recognition rates vary a little according to ϵ, noisy speech recognition results show better performances than standard MFCC results after denoising.

Table 2: Noisy speech recognition results.

SNR (dB)	MFCC (%)	Using 20 basis vectors (%)		
		Before denoising	After denoising ϵ=0.001	After denoising ϵ=0.01
20	91.5	83.6	94.1	92.7
15	78.1	71.5	86.8	88.2
10	54.5	45.1	65.1	76.9
5	23.3	17.6	30.6	45.4

5. CONCLUSION

In this paper, we have obtained efficient speech features using information maximization algorithm of ICA. Many of the ICA features are localized both time and frequency and much similar to Gabor filters. And these speech features efficiently encode the input speech signals than discrete Fourier basis functions. The ICA features were also applied to the automatic speech recognition systems and demonstrated better recognition performance than the standard MFCCs features. Trained ICA speech features were applied for the removal of Gaussian noise in MAP estimator. Speech signals corrupted by additive white Gaussian noise can be recovered with much better SNR values after the MAP denoising process. Finally, denoised speech features showed better recognition results than the standard MFCCs features.

6. REFERENCES

[1] A.J. Bell and T.J. Sejnowski, "The "Independent Components" of natural scenes are edge filters," *Vision research*, vol. 37,(23), pp. 3327-3338, 1997.

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

[3] J.H. Lee, H.Y. Jung, T.W. Lee and S.Y. Lee, "Speech feature extraction using independent component analysis," *accepted in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000.

[4] T.W. Lee, *Independent component analysis - Theory and applications.* Boston: Kluwer Academic Publishers, 1998.

[5] S. Amari, A. Cichocki and H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Indormation Processing Systems*, vol. 8, pp. 757-763, 1996.

[6] M.S. Bartlett, H.M. Lades and T.J. Sejnowski, "Independent component representations for face recognition," *in Proc. SPIE Symposium on Electronic Imaging: Science and Technology; Con ference on Human Vision and Electronic Imaging III*, San Jose, California, January 1998.

[7] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17,(1), pp. 25-46, 1997.

[8] B.A. Olshausen and D.J. Field, "Emergence of simple cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607-609, 1996.

[9] A. Hyvärinen, "Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11(7), pp. 1739-1768, 1999.