LEARNING STATISTICALLY EFFICIENT FEATURES FOR SPEAKER RECOGNITION

Gil-Jin Jang[1], *Te-Won Lee*[2], *and Yung-Hwan Oh*[1]

[1]Spoken Language Laboratory, Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
jangbal@bulsai.kaist.ac.kr, yhoh@cs.kaist.ac.kr

[2]Howard Hughes Medical Institute, Computational Neurobiology Laboratory
The Salk Institute, La Jolla, California 92037, USA
and Institute for Neural Computation, University of California, San Diego
La Jolla, California 92093, USA
tewon@inc.ucsd.edu

ABSTRACT

We apply independent component analysis (ICA) for extracting an optimal basis to the problem of finding efficient features for a speaker. The basis functions learned by the algorithm are oriented and localized in both space and frequency, bearing a resemblance to Gabor functions. The speech segments are assumed to be generated by a linear combination of the basis functions, thus the distribution of speech segments of a speaker is modeled by a basis, which is calculated so that each component should be independent upon others on the given training data. The speaker distribution is modeled by the basis functions. To asses the efficiency of the basis functions, we performed speaker classification experiments and compared our results with the conventional Fourier-basis. Our results show that the proposed method is more efficient than the conventional Fourier-based features, in that they can obtain a higher classification rate.

1. INTRODUCTION

Currently, one of the main focus in speaker recognition research is based in finding efficient features for speech signals, and so far the standard Fourier basis has taken the leading role. In Fourier basis speech signals are decomposed into a superposition of a finite number of sinusoids and their coefficients are used for speaker recognition. However, it is not necessarily able to express the domain's statistical structure, but assumes that all the signals are infinitely stationary and that the probabilities of the basis functions are all equal. Independent component analysis (ICA) [1, 2] has suggested statistical ways of constructing basis for encoding patterns, including images [3, 4] and natural sounds [5]. ICA has been also shown a highly effective in extracting the features from the given set of observed speech signals [6], by reflecting the statistical structure of the observed signals. Recent work showed that the ICA features of speech signals are localized in both time and frequency [5, 6], while the conventional Fourier basis is localized only in frequency. Although the ICA features behave like short-time Fourier basis, they are different in that they are asymmetric in time.

In this paper, we focus on the difference of the statistical structures among the speakers. The ICA filters maximize the amount of information in the transformed domain, so that the adapted individual basis functions obtained by ICA can model the distribution of the individual speaker. In estimating the probability density functions for the sources of the speech basis, previous work adopted a Laplacian prior [6]. However, since we do not want to impose a certain density on the sources we employ the generalized form of Gaussian functions or also called the generalized exponential power function [7], which can model the wide range of distributions. We compare the ICA-based features with the Fourier and PCA by the speaker classification experiments on 20 speakers from the TIMIT database. The source coefficients for each basis function are modeled by the generalized Gaussian density, then the speaker is classified by the one which has the highest likelihood given the all the basis functions for each class. From the results we prove that the proposed features are more effective in describing the statistical structures of speakers.

2. LEARNING THE ICA SPEAKER BASIS

For the observed speech segment with length N, denoting it as $N \times 1$ column vector $\mathbf{x}$, we assume that it can be represented as a linear combination of the N unknown sources s_i such that

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i, \qquad (1)$$

where $\mathbf{s}$ is the source vector constructed by s_i's, $\mathbf{A}$ a scalar square matrix and the column vector $\mathbf{a}_i$'s of $\mathbf{A}$ are the basis functions. Note that $\mathbf{A}$ have to be square and full rank to be a complete basis. $\mathbf{A}$ represents the basis functions generating the observed segments of speech signal in the real world whereas $\mathbf{W} = \mathbf{A}^{-1}$ refers to filters that transform the segments into activations or source coefficients $\mathbf{s} = \mathbf{W}\mathbf{x}$.

For Fourier basis, each $\mathbf{a}_i$ is a complex sinusoid with its own frequency and unit magnitude, resulting in mutual exclusion —orthonormality— with the other sinusoids. ICA basis is different in that the basis functions are real and not necessarily orthonormal, and the sources are statistically independent. The ICA basis reflects the statistical information of the short-time speech segments from the training data, because ICA is formulated as one of density estimation of the sources [1]. We use the infomax learning rule for updating the basis functions:

$$\Delta\mathbf{A} \propto \mathbf{A}(\mathbf{I} - \varphi(\mathbf{s})\mathbf{s}^T)\,. \qquad (2)$$

where the vector $\varphi(\mathbf{s})$ is a function of the prior and is defined by $\varphi(\mathbf{s}) = -\frac{\partial \log p(\mathbf{s})}{\partial \mathbf{s}}$. For the density model for sources, $p(s_i)$, We use a flexible prior known as generalized Gaussian [7] that can change the overall shape of the density functions.

2.1. The Generalized Gaussian Distributions

The generalized Gaussian models density functions that are peaked and symmetric at the mean, with the varying degree of normality in the following general form [7, 8]:

$$p(x|\mu,\sigma,q) = \frac{\omega(q)}{\sigma}\exp\left[-c(q)\left|\frac{x-\mu}{\sigma}\right|^{q}\right]\,, \qquad (3)$$

where $\mu = E[x]$, $\sigma = \sqrt{E[(x-\mu)^2]}$, $c(q) = \left[\frac{\Gamma[3/q]}{\Gamma[1/q]}\right]^{q/2}$, and $\omega(q) = \frac{\Gamma[3/q]^{1/2}}{(2/q)\Gamma[1/q]^{3/2}}$ [1]. The exponent q controls the distribution's deviation from normality. The Gaussian, Laplacian, and strong Laplacian —speech signals— distributions are modeled by putting $q = 2$, $q = 1$, and $q < 1$ respectively. Note that the distribution approaches delta function as q goes to 0. Parameter q can also be converted to the standard kurtosis measure $K = E[(x-\mu)^4/\sigma^4] - 3$:

$$K = \frac{\Gamma[5/q]\Gamma[1/q]}{\Gamma[3/q]^2} - 3. \qquad (4)$$

As K increases the distribution gets sparser because in the highly peaked distribution almost all the datapoints are close to zero and the few non-zero coefficients are scattered sparsely.

2.2. The Generalized Gaussian ICA

For the purposes of finding the basis functions in ICA, zero mean and unit variance is assumed. Because the components are statistically independent, the likelihood of the source vector is factorized in the generalized Gaussian form as

$$p(\mathbf{s}|\mathbf{q}) = \prod_{i}^{N} \omega(q_i)\exp\left[-c(q_i)|s_i|^{q_i}\right]\,, \qquad (5)$$

where $\mathbf{q} = [q_1 q_2 \dots q_N]^T$, and $\{q_i\}$'s are the exponents of the source distributions. In equation 2, each component of the gradient vector $\varphi(\mathbf{s})$ is derived from $p(\mathbf{s}|\mathbf{q})$ as

$$\varphi_i(s_i) = -\eta|s_i|^{q-1} qc\sigma_i^{-q}\,, \qquad (6)$$

where $\eta = \mathrm{sign}(s_i)$, c and σ are defined in equation 3. Detailed derivations of the density function and the learning rule are given in [7]. Varying the parameters q_i by updating them periodically during the adaptation process enables $p(s_i)$ to match the distribution of the estimated sources exactly. Gradient ascent is used to estimate the parameters that maximize the log Likelihood. Figure 1 shows the obtained bases of 4 speakers —2 male and 2 female— by generalized Gaussian learning rule. The data are from TIMIT database. They have quite different shape in the locality of time and frequency.

3. SUPERVISED CLASSIFICATION OF SPEAKERS

The performance of the proposed features of speakers, we performed an experiment of supervised classification of speakers. We use the generalized mixture model [9] on estimating the density functions of the coefficients of each basis.

3.1. The Generalized Mixture Model Using ICA

The likelihood of the speech data for a given model is calculated by a generalized Gaussian mixture model. A mixture density is defined as [10]:

$$p(\mathbf{x}_n|\Theta) = \sum_{k=1}^{K} p(\mathbf{x}_n|C_k,\theta_k)p(C_k), \qquad (7)$$

where $\Theta = (\theta_1, \cdots, \theta_K)$ are the unknown parameters ($\mathbf{A}_k$, $\mathbf{b}_k$, $\mathbf{q}_k$) for the component densities $p(\mathbf{x}_n|C_k,\theta_k)$. For the present model, the class log likelihood is given by the log likelihood for the standard ICA model:

$$\log p(\mathbf{x}_n|\theta_k, C_k) = \log p(\mathbf{s}_k) - \log|\det \mathbf{A}_k|, \qquad (8)$$

[1] For notational compactness, we define the parameters c and ω in the different forms with [7, 8].

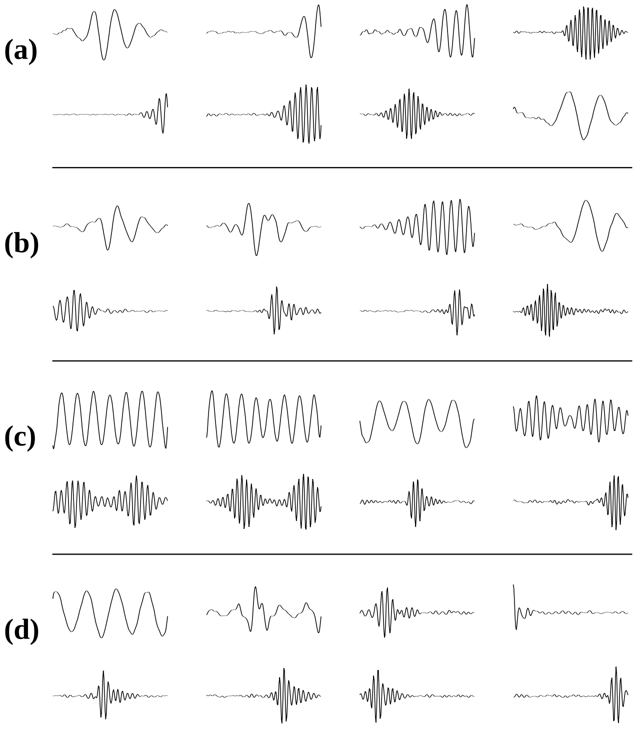

Fig. 1. Example plots of learned ICA basis functions. (a), (b): male speakers, (c), (d): female speakers. Each basis function is up-sampled by 5 to remove artifacts from sample aliasing. Only 8 basis functions are shown among the 64. They are obtained by the generalized Gaussian ICA learning algorithm from the 64-sample speech segments from TIMIT database.

where $\mathbf{s}_k = \mathbf{A}_k^{-1}(\mathbf{x}_n - \mathbf{b}_k)$, the coefficients of the basis, and $\mathbf{b}_k$ is the mean vector of the coefficients. For Fourier basis, the linear transformation $\mathbf{A}_k$ is an orthonormal set of sinusoidal functions. Thus $\log|\det \mathbf{A}_k|$ is zero because $\mathbf{A}_k\mathbf{A}_k^T = \mathbf{I}$.

The classification is done by processing each data instance with the learned parameters $\mathbf{q}_k$, $\mathbf{A}_k$ and $\mathbf{b}_k$. The probability of the class $p(C_k|\mathbf{x}_n, \theta_k)$ is computed and the corresponding instance label is compared to the highest class probability. The priori probabilities of speakers are assumed to be equal, that is, in equation 7, $p(C_k) = 1$ for all k, because the models are trained in a supervised manner. The speaker is classified by the maximum likelihood.

3.2. Learning Data and Testing Data

From the TIMIT databast, 20 speakers are randomly selected. 7 sentences for each speaker are selected from the SX (phonetically-compact) and the SA (dialect) set, 4 of them used for training the each basis, 3 of them for testing. Training and testing sets have no intersection. Because each data are labeled with a speaker ID, we learn each speaker basis with only that speaker's data, in supervised manner. We down-sampled the originally 16kHz-sampled data to 8kHz and applied pre-emphasis with $1 - 0.95z^{-1}$, to complement the energy decrease in the high bands of human speech. Those processes reduce the redundancy and prevent low-frequency component from dominating the gradient. The learning data $\mathbf{x}$ were constructed from the speech data segmented in 64 samples (8ms) blocks. The adaptation started from the random 64×64 square matrix $\mathbf{A}$, and the gradient of basis functions was computed on a block of 1000 waveform segments. The parameter q_i for each $p(s_i)$ was updated every 10 gradient steps, and the learning rate was gradually decreased from 0.001 to 0.0001 as the iterations went on. The parameters q_i are updated periodically during the adaptation process.

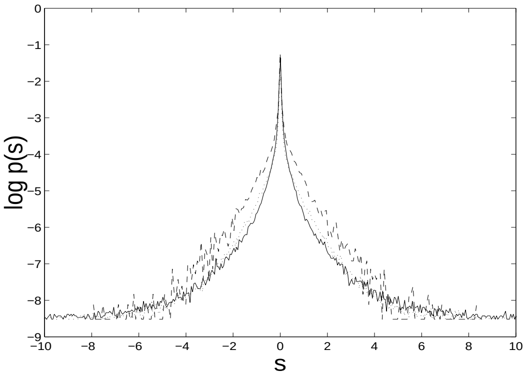

Fig. 2. Distributions of basis function coefficients for ICA, PCA, and Fourier basis. The solid line is ICA, dotted PCA, and the dash-dotted Fourier coefficients. The data are from the male speaker 'mgrl0' in TIMIT. Note that the y-axis is log scale.

To compare the performance of the proposed features with conventional method, we trained the generalized Gaussian mixture model by the real part of the Fourier transformation of the given training data. Figure 2 compares the log-scaled histograms of each Fourier, PCA, ICA coefficients. ICA coefficients have higher kurtosis than the other PCA and Fourier basis. In figure 3 the dependency of the coefficients decreases significantly from Fourier and PCA bases.

4. EXPERIMENTAL RESULTS

We report and compare the rate of correct classification rate and the average kurtosis of each basis in table 1. The kurtosis is derived from the estimated exponent q_i by equation 4 and averaged by geometric mean, because large value of kurtosis possibly dominates the small values. In the speaker

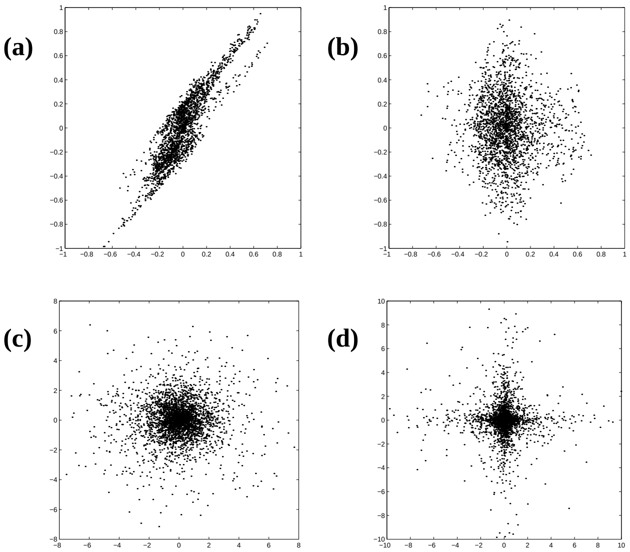

Fig. 3. 2-dimensional plots for the coefficients of each basis, first versus second coefficient: (a), (b) Fourier [2, 3] and [2, 20]; (c) PCA [1, 2]; (d) ICA [1, 2]. (a) shows that Fourier basis has high correlation between adjacent coefficients.

recognition experiments, the individual ICA basis is the most effective, both in sparseness (kurtosis) and the classification rate. Using ICA basis the sparseness increased and thus the distributions of the coefficients became more apparent to classify as the increased classification showed.

5. CONCLUSION

We applied ICA to speech signals from individual speakers to extract a set of optimal basis functions. The basis functions were adapted using the generalized Gaussian ICA model resulting in basis functions and source coefficient statistics that were characteristical features for the individual speaker. Most basis functions were localized in time and frequency resembling Gabor-like wavelet filters. The corresponding source coefficients were extremely sparse resulting in efficient codes. The generalized Gaussian ICA model is embedded into a mixture model allowing classification of the individual speakers based on the basis functions models for each speaker class. Our initial recognition rates suggest superior performance compared to the Fourier or PCA based method. This can now serve as a baseline to further investigate and optimize the classification procedure. Then, we plan to compare our results to state of the art speaker recognition systems.

Table 1. Correct Classification Rates and the mean value of Kurtosis for each basis

	Fourier	PCA	ICA
Classification Rate	82.2%	84.1%	87.5 %
Kurtosis	181.2	239.1	248.7

Acknowledgements

T.-W. Lee was partially supported by the Digital Media Innovation Program (DiMI) and the NSF CCR-9902961 grant. The authors would like to thank Jee Hea An for proofreading the paper.

6. REFERENCES

[1] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1004–1034, 1995.

[2] P. Comon, “Independent component analysis, A new concept?,” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[3] A. J. Bell and T. J. Sejnowski, “The “independent components” of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[4] 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.

[5] A. J. Bell and T. J. Sejnowski, “Learning the higher-order structures of a natural sound,” *Network: Computation in Neural Systems*, pp. 261–266, July 1996.

[6] J.-H. Lee, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, “Speech feature extraction using independent component analysis,” in *Proc. ICASSP*, Istanbul, Turkey, June 2000, vol. 3, pp. 1631–1634.

[7] M. S. Lewicki, “A flexible prior for independent component analysis,” *Neural Computation*, 2000.

[8] G. Box and G. Tiao, *Baysian Inference in Statistical Analysis*, John Wiley and Sons, 1973.

[9] T.-W. Lee and M. S. Lewicki, “The generalized gaussian mixture model using ICA,” in *International Workshop on Independent Component Analysis (ICA’00)*, Helsinki, June 2000, pp. 239–244.

[10] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, John Wiley and Sons, Inc., 1973.