THE STATISTICAL STRUCTURES OF MALE AND FEMALE SPEECH SIGNALS

Te-Won Lee

Computational Neurobiology Laboratory,
The Salk Institute, La Jolla, CA 92037, USA
and Institute for Neural Computation
University of California, San Diego,
La Jolla, CA 92093, USA
tewon@salk.edu

Gil-Jin Jang

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

ABSTRACT

The goal of this paper is to learn or adapt statistical features of gender specific speech signals. The adaptation is performed by finding basis functions that encode the speech signal such that the resulting coefficients are statistically independent and the information redundancy is minimized. We use a flexible independent component analysis (ICA) algorithm to adapt the basis functions as well as the source coefficients for male and female speakers respectively. The learned features show significant differences in frequency and time span. Our results suggest that the male speech features can be described by Gabor-like wavelet filters whereas the female speech signal has a much longer time span. We present a detailed time-frequency analysis strongly suggesting that those features can be used to qualify and quantify gender-specific speech signal differences.

1. INTRODUCTION

The efficient encoding of speech signals is a crucial step in the most speech recognition systems. Most commonly used features are based on statistics of 2nd order such as the standard Fourier transformation that leads to the cepstral representation of the speech signal. Although this representation may be computationally simple and its model appealing it would be more interesting to adapt or learn features that are natural and represented in the structure of the speech signals.

Recently, independent component analysis (ICA) [1, 2, 3] has been shown highly effective in encoding patterns, including images and speech signals [4, 5]. ICA assumes that the speech signal can be decomposed into basis functions and coefficients. The basis functions can be adapted by a standard ICA learning rule. This technique was employed in [6] to learn the basis of speech signals in general and to show that the ICA features (basis functions) of speech signals are localized in both time and frequency, while the conventional Fourier bases are localized only in frequency. The learned features can be used in standard pattern recognition systems to achieve speech recognition performance that are comparable to current MFCC features-based recognition systems. The ICA source coefficients are usually assumed to have a sparse distribution [4] resulting then in statistically efficient codes. In many ICA algorithms, this prior density model is assumed fixed [4, 5]. Here, we use a more flexible prior allowing the source coefficient statistics to be inferred from the data.

In this paper, we focus on the difference of the statistical structures of male and female speakers. Although the ICA features behave like short-time Fourier bases; similar to Gabor features. They are however different in the fact that they are asymmetric in time. Encoding and comparing both the spectral and temporal properties of male and female speakers present a statistical explanation and understanding of their difference. We used a subset of the TIMIT database to obtain the basis functions of male and female speech signals using the generalized Gaussian densities [7] to model the distribution of the source coefficients. We compare the statistical structures of male and female speech signals in terms of their time and frequency span. Furthermore, we analyze their coding efficiency in terms of the sparseness and the independence assumption of the source coefficients.

2. THE ICA ALGORITHM

We assume an unknown source vector $\mathbf{s}$ with statistically independent components s_i. The observed data $\mathbf{x}$ is represented as a linear combination of s_i such that

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

where $\mathbf{A}$ is a scalar square matrix and the column vector $\mathbf{a}_i$'s of $\mathbf{A}$ are the basis functions. $\mathbf{A}$ represents the basis functions generating the observed segments of the speech signal in the real world whereas $\mathbf{W} = \mathbf{A}^{-1}$ refers to the ICA filters that transform the segments into activations or source coefficients $\mathbf{s} = \mathbf{W}\mathbf{x}$. The objective of ICA is to infer both the unknown sources s_i and the unknown basis functions $\mathbf{A}$ from the data signal, and it is formulated as one of density estimation of the sources [1, 2]. In our experiments, 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 defined by $\varphi(\mathbf{s}) = -\frac{\partial \log p(\mathbf{s})}{\partial \mathbf{s}}$. $\Delta\mathbf{A}$ will converge to zero matrix when the basis functions are completely adapted. We use a generalized Gaussian $p(s_i) \propto \exp(-|s_i|^{q_i})$, and derive each component of $\varphi(\mathbf{s})$ as

$$\varphi_i(s_i) = -\eta |s_i|^{q-1} q c \sigma_i^{-q}\,, \qquad (3)$$

where $\eta = \text{sign}(s_i)$, $c = [\Gamma(3/q)/\Gamma(1/q)]^{q/2}$, and $\sigma_i = \sqrt{E[s_i^2]}$. 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. Note that $\varphi_i(s_i)$ was fixed in previous works [4, 5].

3. ADAPTING BASIS FUNCTIONS FOR SPEECH SIGNALS

To learn the basis functions of male and female speeches, 7 male speakers and 7 female speakers were randomly chosen from the 462 speakers of the TIMIT database. 4 sentences were selected from the SX (phonetically-compact) set for each speaker, and the average duration of one sentence is about 3 seconds. 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 training data $\mathbf{x}$ were constructed from the speech data segmented in 64 samples (8ms) starting from every samples. The adaptation started from the 64×64 PCA basis functions $\mathbf{A}$, 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.

We obtained two different set of basis functions from the separated male and female data. Subsets of 32 learned basis functions are shown in figure 1. They look like short-time Fourier bases, but are different in that they are asymmetric in time. They often show Gabor-like filter characteristics, having a peak rising and decaying slowly. The basis functions of male speakers usually have one Gabor peak, but those of female speakers have a few peaks, generally two, or cover all time span like Fourier basis.

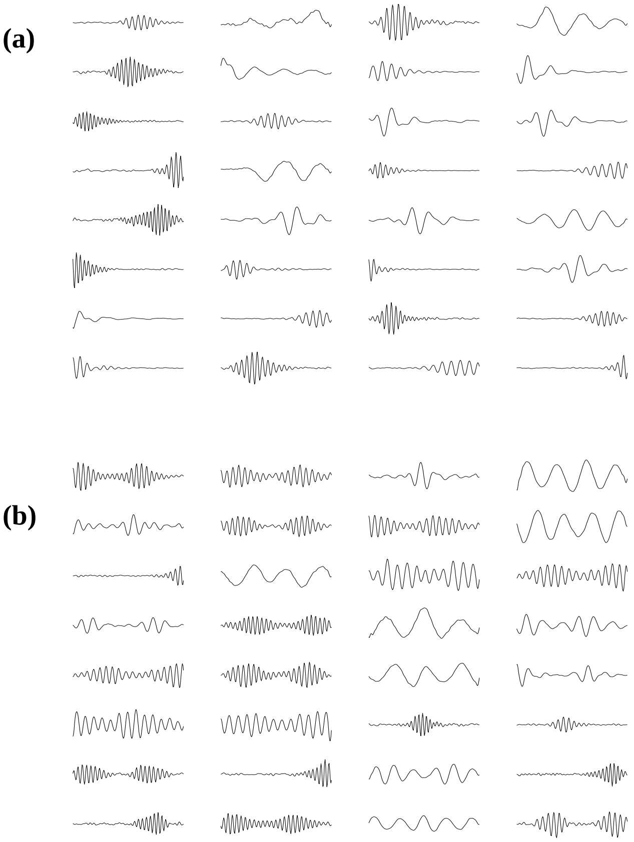

Fig. 1. Learned ICA basis functions of (a) male and (b) female speakers. Both are obtained by the generalized Gaussian ICA learning algorithm from speech segments of 64 samples, and 32 basis functions are selected out of the 64. Each basis function is up-sampled by 5 to remove artifacts from sample aliasing.

4. COMPARISON OF MALE AND FEMALE BASIS FUNCTIONS

To compare the structures of different sets of basis for time-varying signals we analyzed the ways the basis functions tile the time-frequency space. In figure 1, almost all of the male basis are highly localized in time, i.e. some basis functions are active only over a brief time period. In contrast, the female basis functions are much less localized in time, but they tend to cover a broader frequency span.

A standard Fourier basis represents signals by a superposition of exclusive sinusoids. The basis functions are localized uniformly in frequency, but not in time. In figure 2-(a), each rectangle corresponds a Fourier basis, and carries

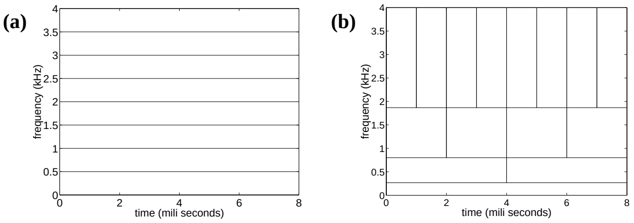

Fig. 2. Plot of the time and frequency spans of conventional bases. (a) The basis functions in a Fourier basis are localized only in time. (b) Wavelet basis functions are localized in both time and frequency. Both sets of basis functions have no intersection, with carrying equal amount of area for each basis.

the same amount of information on the all time span, with no intersection on the frequency. In figure 2-(b), a wavelet basis is shown to be composed of basis functions that are localized in both time and frequency. The two basis sets contain the same number of functions, so while a wavelet basis provides improved resolution in time, it necessarily sacrifices some resolution in frequency. For purpose of coding, which tiling is best depends on the structure of the signals being analyzed.

4.1. Time Frequency Analysis

Because the majority of learned basis functions are localized in both time and frequency, it is possible to plot how they cover the time-frequency space. Figure 3 shows that the learned features for male and female speech signals cover time and frequency space in a manner similar to a wavelet representation. For PCA, the basis functions are almost identical, therefore only the male basis functions are presented. Each ellipse refers to one basis function with the extent of its coverage in time-frequency space. We define the time span and the frequency span as follows: the time span —the horizontal width of the ellipse— is the temporal extent required to cover 95% of the signal power, and the frequency span —the vertical height of the ellipse— is the width of the largest spectral peak at 5% maximum. Both represent the extent that the basis covers in time and frequency. This conveys accurately the extent of each basis in time-frequency space, although with a considerable degree of overlap. In figure 3 (a), PCA basis, the frequency bandwidth is relatively constant across frequency. In (b) and (c), it gradually decreases in temporal bandwidth and increases in frequency with increasing frequency, and the male basis functions have a higher tendency towards smaller temporal bandwidth. Note that they gradually change like critical band frequencies, rather than having discrete octave increases in bandwidth, which is common for many types of wavelets.

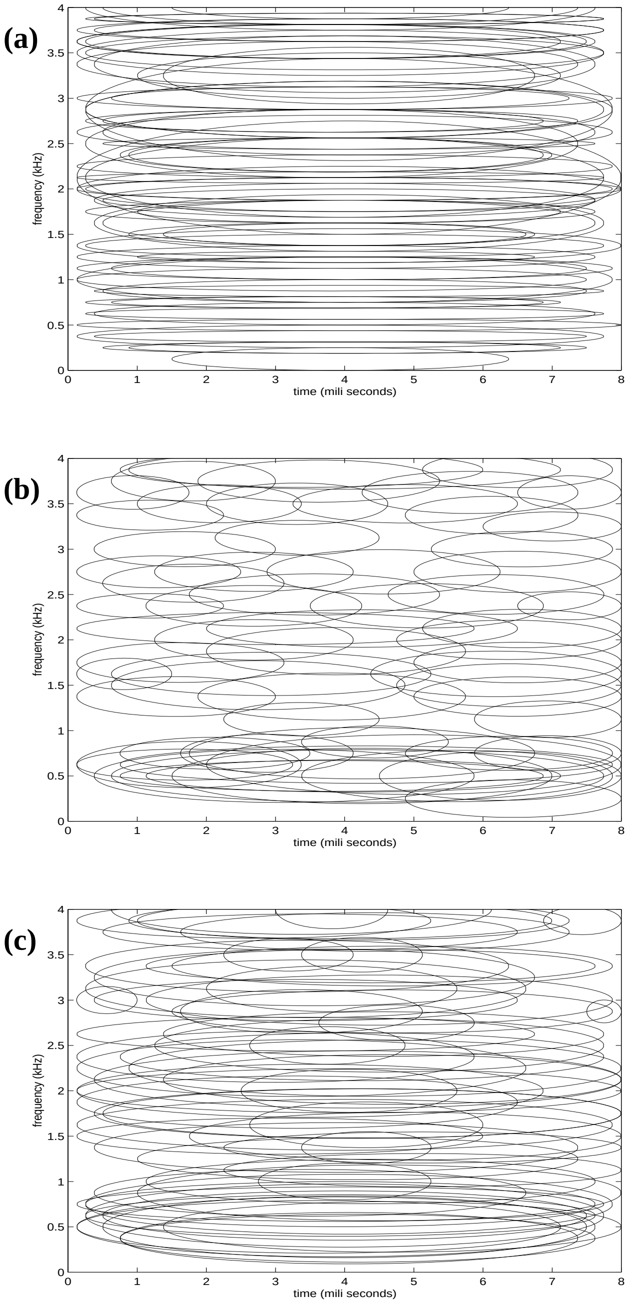

Fig. 3. Time-frequency analysis of male and female basis. (a) PCA basis of male speakers. (b) ICA basis of male speakers. (c) ICA basis of female speakers. Because the two PCA basis of male and female speakers are almost identical, only male basis are shown. Each ellipse indicates the range to cover 95% of the signal power in time and the width at 5% of maximum in frequency.

4.2. Statistical Analysis

The locality of a basis function is closely related to the resolution in each domain of time and frequency — the time

Table 1. Localities and sparseness for Fourier, PCA male and female, ICA male and female bases.

	Time span	Freq. span	kurtosis
Fourier	8.0 ms	62.5 Hz	21.5
PCA-male	7.1 ms	100.3 Hz	19.4
PCA-female	7.2 ms	107.2 Hz	26.2
ICA-male	3.4 ms	175.5 Hz	28.8
ICA-female	5.6 ms	167.0 Hz	36.6

span is inversely proportional to the time resolution, and the frequency span to the frequency resolution. Table 1 shows the spans of 5 bases: Fourier, PCA male and female, and ICA male and female. Both PCA and ICA bases are obtained from the same speech data. Fourier basis have a full range of the time resolution in its analysis length, but their frequency span is fixed. In ICA bases, the spans vary in male and female. Although the frequency spans are almost the same, the male basis have comparatively shorter time span. Note from the table 1 that the time span of male and female are quite similar in case of the PCA bases.

Another crucial property is the amount of information conveyed by the source coefficients of the basis, which are proportional to the sparseness of the sources. We approximately measured the source sparseness with the standard kurtosis measure, defined by $K(s) = E[(s - \bar{s})^4/\sigma_s^4] - 3$. Kurtosis is a measure of peakness, or super-Gaussianity. In the peaked, super-Gaussian distribution, almost all the datapoints are close to zero and the only few non-zero coefficients are scattered sparsely. Coding efficiency and statistical independency increase as the source coefficients become more sparse. The last column of the table 1 shows the averaged kurtosis of each basis calculated by the geometric mean. All the values are computed by the male and female data used for training the basis. For Fourier basis, all the male and female data are used. The obtained values indicate that the PCA basis for male and female speech are very similar whereas in case of the ICA bases, there are significant differences in male and female time-frequency plots of the data. This is mostly due to the exploitation of the higher-order structure in the speech signal using ICA.

5. CONCLUSION

We analyzed a set of basis functions obtained for male and female speech signals. The basis functions as well as the source coefficient statistics are key factors that capture the statistical structure of speech signals. The appropriate learning rule for this task was ICA with a flexible prior. Our results suggest that the learned coefficients are extremely sparse, making them useful as statistically efficient codes for many applications. Again, we note that this sparseness was not enforced by the algorithm but was a mere result of the independence assumption that was required to reduce the redundancy in the data. The basis functions for male and female speech signals were quite different. The time-frequency analysis indicate their difference clearly suggesting that the Fourier transformation is more likely to capture the statistics of the female speech than the male speech signal. Male speech features are more reminiscent of Gabor-like wavelet features. We have extended our results in analyzing speaker specific basis functions for speaker coding and recognition. Our initial results suggest a simple and maybe natural framework for speaker recognition.

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] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on S.P.*, vol. 45, no. 2, pp. 424–444, 1996.

[4] 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.

[5] 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.

[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.