Learning self-organized topology-preserving complex speech features at primary auditory cortex

Taesu Kim and Soo-Young Lee

Brain Science Research Center and Department of BioSystems
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

Abstract

By applying independent component analysis (ICA) algorithm to auditory signals a computational model was developed for the speech feature extraction at the primary auditory cortex. Unlike the other ICA-based features with simple frequency selectivity at the basilar membrane and inner-hair-cells the learnt features represent complex signal characteristics at the primary auditory cortex such as onset/offset and frequency modulation in time. Also, the topology is preserved with the help of neighborhood coupling during the self-organization. The extracted complex features demonstrated good performance for the robust discrimination of speech phonemes.

Keywords: Independent component analysis, Auditory cortex , Complex speech features, Topology-preserving self-organization, Neural coding

1. Introduction

The computational models of human auditory nerve systems had attracted many attentions from both neuroscience and speech recognition communities. Although simple frequency-selective features are extracted at the basilar membrane and inner-hair-cells (IHCs) in the cochlea, physiological experiments prove that more complex features are extracted along the auditory pathway. Especially, some neurons at the primary auditory cortex respond to specific time-frequency signals such as onset/offset and frequency modulation in time. It is believed that this signal coding (or feature extraction) mechanism follows the information theory.

Recently, based on an information-theoretic theory of self-organization, i.e., independent component analysis (ICA), several researchers had come up with computational models of feature extraction from natural and speech sounds [1-4]. These self-organized speech features resemble the frequency-selectivity of the IHC, but do not show the complex time-dependent features at the primary auditory cortex.

In this paper, we report an ICA-based computational model for the complex speech features at the auditory cortex. The extracted features from human speeches provide the topology-preserving mapping of speech signals. It is also demonstrated that the extracted speech features may be used for human speech perception.

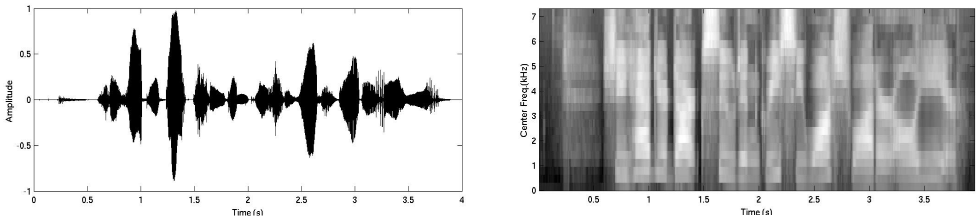
2. Computational Model of Auditory Nerve Systems

The feature extraction at cochlea is relatively well-understood [5], and we are interested in the mapping from the cochlear features to those at the auditory cortex only. Therefore, we use the existing models of cochlea to generate speech features at the inner-hair-cells, which is the input stage of the ICA-based mapping network.

We first modeled the outer and middle ear resonances as a simple linear high-pass filter as

$$y(t) = x(t) - 0.9x(t-1)\,, \tag{1}$$

where *x(t)* is the speech signal at time step *t* and *y(t)* is the filtered output signal, which goes to inner ear and vibrates the tympanic membrane at the cochlea. Secondly, for the cochlear filters, a Mel-frequency filterbank is used for computationally efficiency. Actually the Mel-frequency filterbank is widely used in automatic speech recognition [6]. Finally, to model the logarithmic sensitivity of the loudness, the outputs of the Mel-scaled filterbank go through a logarithmic nonlinearity [5]. The results may be considered as average firing rates in the IHCs, which stimulate the higher auditory pathway. Figure 1 shows an example of the speech signal and corresponding average firing rates in time and frequency domain.



(a) (b)

Figure 1: (a) Time-domain acoustic signal. (b) Average firing rate in the auditory nerve (IHC)

In both vision and auditory nerve systems the feature coding may be understood in the framework of the sparse coding or the ICA for super-Gaussian sources [7]. Also, an experiment on the dynamic functional connectivity in an auditory cortex shows the sparse connectivity [8]. Therefore, to get complex features at auditory cortex, we had applied an ICA algorithm for the IHC signals in Figure 1(b). It is worthy noting that the other ICA-based speech features were extracted from the time-domain speech signals in Figure 1(a), and resulted in frequency selectivity only. [1-4]

The basic idea of ICA is to find the hidden random variables **s**, when the observed data (random variables) **x** is given as a linear superposition of **s** as

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{s} + \mathbf{n}\,, \tag{2}$$

where **A** is an M × N mixing matrix. [7][9] By maximizing log-likelihood of **x** with a given **A**, the learning rule can be derived as following:

$$\Delta\mathbf{A} \propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log P(\mathbf{x} \mid \mathbf{A}) = -\mathbf{A}(\mathbf{I} - \varphi(\hat{\mathbf{s}}) \cdot \hat{\mathbf{s}}^T), \tag{3}$$

where $\varphi(\hat{s}_i) = -\partial \log P(\hat{s}_i) / \partial s_i$ is the score function, **I** is the identity matrix, and $\mathbf{AA}^{\mathrm{T}}$ is used for the faster convergence [10]. $\hat{\mathbf{s}} = [\hat{s}_1\ \hat{s}_2\ \Lambda\ \hat{s}_N]^T$ is inferred hidden variable, which can be obtained by finding the maximum a *posteriori* value of **s**:

$$\hat{\mathbf{s}} = \max_{\mathbf{s}} P(\mathbf{s} \mid \mathbf{x}, \mathbf{A}) = \max_{\mathbf{s}} P(\mathbf{x} \mid \mathbf{A}, \mathbf{s})\, P(\mathbf{s}) . \tag{4}$$

When the mixing matrix is rectangular and the additive noise does not exist, the solution for **s** can be found as $\hat{\mathbf{s}} = \mathbf{A}^{-1} \cdot \mathbf{x}$.

To implement topologically-preserving mapping from the IHC signals to those of auditory cortex, we also introduce a neighborhood function as [11]

$$h_{ij} = \exp[-\frac{(i-j)^2}{2\sigma^2}] . \tag{5}$$

where σ is the width of the neighborhood. Then complex cell output, $\mathbf{y} = [y_1\ y_2\ \Lambda\ y_i]^T$, is defined as

$$y_i = \sum_j h_{ij} \cdot \hat{s}_j^{\,2} . \tag{6}$$

Then, the modified ICA learning rule becomes

$$\Delta\mathbf{A} \propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log P(\mathbf{y} \mid \mathbf{A}) = -\mathbf{A}(\mathbf{I} - \psi(\hat{\mathbf{s}}) \cdot \hat{\mathbf{s}}^T), \tag{7}$$

where

$$\psi(\hat{s}_j) = \hat{s}_j \cdot \sum_i h_{ij} \cdot \varphi(y_i) , \tag{8}$$

$$\varphi(y_i) = -\partial \log P(y_i) / \partial y_i . \tag{9}$$

For speech signal the probability density function of y_i can be assumed as supper-Gaussian. Thus, we used the following probability density function

$$p(y_i) = \alpha \cdot \exp(-\beta\, y_i^{1/2}) . \tag{10}$$

Figure 2 shows the overall procedure of the method.

3. Speech Features at Auditory Cortex

In order to adapt the ICA network for self-organization, we used the TIMIT continuous speech corpus as the training data. It consists of 630 speakers, 438 males and 192 females. Ten sentences are pronounced by each speaker, and sampling frequency is 16kHz. To calculate average firing rates in the inner-hair-cells using the Mel-frequency filterbank, a hamming window with 30ms frame length and 10ms shift is used. For computational simplicity, 23 Mel-scaled filters are selected. The time duration for basic speech signal is set to 5 frames, i.e., 50ms. Thus, the dimension of the data is 23 × 5 = 115. To remove the effect of the average component of the data, the local mean was subtracted. To remove noise and reduce computational complexity, the data dimension is reduced to 81 by principle component analysis (PCA). Therefore, the input dimension of ICA network is 81. For self-organizing topology-preserving mapping, we determine the neighborhood in an 8 × 8 rectangle. Thus, Eq. (5) was modified to a 2-dimensional neighborhood function. The width, σ, of the neighborhood was set to 2, and then decreased for the fine tuning.

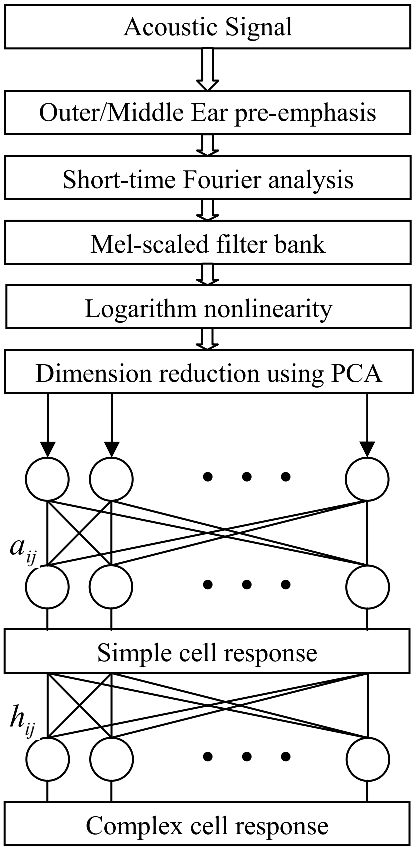


Figure 2: Overall procedure

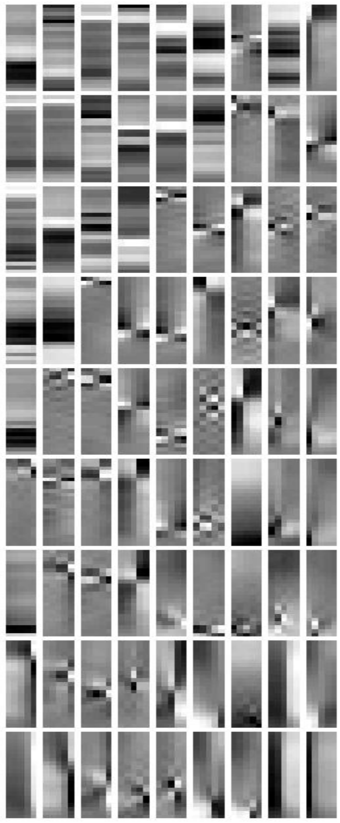
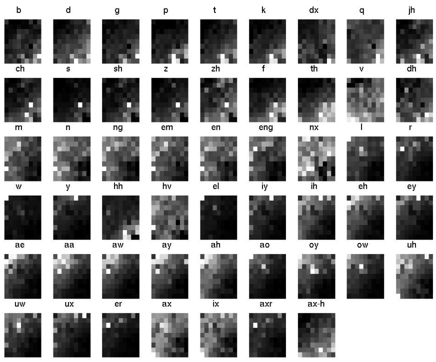
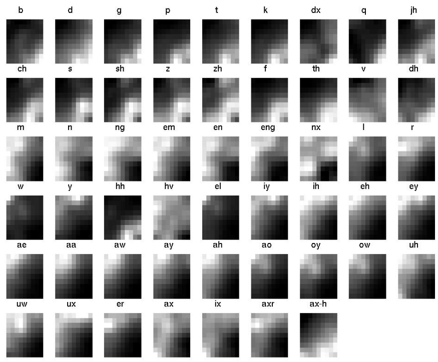


Figure 3: Learnt speech features

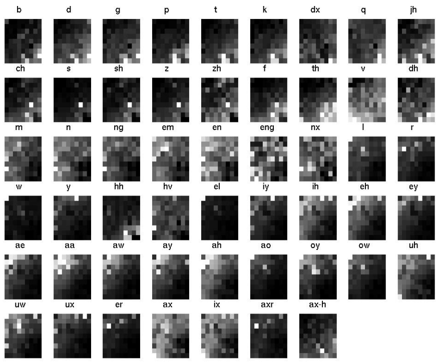
Figure 3 shows the learnt speech features based on self-organizing scheme. They are located topographically in the 2-dimensional map. The horizontal lines at the upper left corner represent time-independent frequency-maintaining components. In this case, more than two frequency components are appeared concurrently. In contrast to horizontal lines, vertical lines are appeared at the lower right corner of the map, and represent noise bursts or signal on- and off-set components. In the middle of the map, lots of frequency-transition are found. They are frequency-rising and frequency-falling components, which can be understood as frequency modulation (FM) components. In fact, human speech and animal sounds share the same three basic elements: steady-state harmonically related frequencies, FMs, and onset/offset bursts. Also, many auditory cortical areas are tonotophically. Therefore, our result fits those biological evidences well.

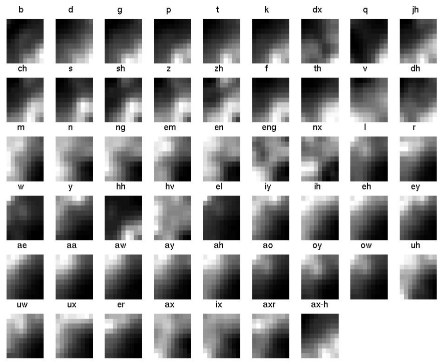
Using the learnt speech features, we investigated the responses to single phonemes, which are complex sounds and also basic elements of speech. The TIMIT corpus provides two data sets, training and testing sets, which consist of different speakers from each other, and also provides the transcription of 52 single phonemes. We calculated average responses to each phoneme in the training and testing sets, respectively. Figures 3(a) and (b) show the result of the training set, and Figures 3(c) and (d) show the result of the testing set. As shown in Figure 3, different phonemes respond in different ways, while different data sets make almost no difference.





(a) (b)





(c) (d)

Figure 4: Average responses of the learnt speech features to 52 phonemes. (a) response in the training set. (b) local-averaged response in the training set. (c) response in the testing set. (d) local-averaged response in the testing set.

4. Conclusion

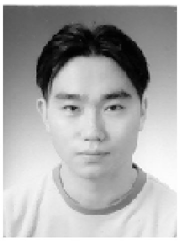
An information-theoretic model based on independent component analysis results in complex speech features such as onset/offset and frequency modulation in time, which are extracted at the primary auditory cortex. Also, similar to the auditory cortex, their topology is preserved. It demonstrates that the signal processing mechanism in human auditory nerve systems follows the information theory for the best efficiency in signal coding and feature extraction. These features are speaker independent, and may be applicable to the automatic speech recognition.

Acknowledgment: This work was supported as a Brain Neuroinformatics Research Program sponsored by Korean Ministry of Science and Technology and also by the Chung Moon Soul BioInformation and BioElectronics Research Center..

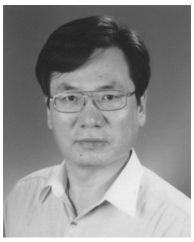
References

[1] A.J. Bell and T.J. Sejnowski, "Learning the higher-order structure of a natural sound," Network : Computation in Neural Systems, Vol. 6, pp. 261–266, 1996.

[2] S.A. Abdallah and M.D. Plumbley, “If the independent components of natural images are edges, what are the independent components of natural sounds?,” in Proc. ICABSS, 2001
[3] M.S. Lewicki, “Efficient coding of natural sounds,” Nature Neuroscience, 2002
[4] J.H. Lee, T.W. Lee, H.Y. Jung, and S.Y. Lee, “On the efficient speech feature extraction based on independent component analysis,” Neural Processing Letters, Vol. 15, pp.235–245, 2002.
[5] W.A. Yost, “Fundamentals of hearing – An introduction,” Academic Press, 2000.
[6] ETSI ES 201 108 speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm compression algorithms, ETSI Standard, 2000.
[7] T.W. Lee, “Independent Component analysis,” Kluwer Academic Publishers, 1998.
[8] R.S. Clement, W. Patrick, J. Rousche, and D.R. Kipke, “Functional connectivity in auditory cortex using chronic,” Neurocomputing Vol. 26–27, pp. 347–354, 1999.
[9] A. Hyvärinen, J. Karhunen, and E. Oja, “Independent component analysis,” John Wiley & Sons, 2001.
[10] S. Amari, A. Cichocki, and H.H. Yang, “A new learning algorithm for blind source separation,” in Proc. NIPS, 1996.
[11] A. Hyvarinen, P.O. Hoyer, and M. Inki, “Topographic independent component analysis,” Neural computation, Vol 13, pp.1527–1558, 2001.
.



Taesu Kim was born in Korea, in 1978. He received the B.S. degree from the Hanyang University in 2001 and the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST) in 2003, both in electrical engineering. He is currently a Ph. D. candidate at the Department of Biosystems, KAIST. His research interests include biologically-motivated information processing, independent component analysis, and noise-robust speech recognition.



Soo-Young Lee received a Ph.D. degree from the Polytechnic Institute of New York in electrophysics in 1984. From 1982 to 1985 he had worked for General Physics Corporation at Columbia, MD, USA. In early 1986 he joined the Korea Advanced Institute of Science and Technology, and now is a Full Professor and Chairman for the newly-established Department of BioSystems. His research interests have resided in artificial auditory systems based on brain information processing mechanism. Especially he have developed an auditory model for noise-robust speech feature extraction, analog and digital chips for speech recognition and adaptive noise canceling, blind signal separation for speech enhancement, and top-down selective attention models for noisy and superimposed pattern recognition. Dr. Lee had served as the President of Asia-Pacific Neural Network Assembly in 2001, and received Leadership Award and Achievement Award from International Neural Network Society in 1994 and 2001, respectively. Dr. Lee is the Editor-in-Chief of a new journal, Neural Information processing – Letters and Reviews.