Feature Extraction and Signal Reconstruction of Air and Bone Conduction Voices by Independent Component Analysis

Tadahiro Azetsu, Eiji Uchino, Ryosuke Kubota, and Noriaki Suetake *

Abstract—**We discuss in this paper the feature extraction and signal reconstruction of air and bone conduction voices in the time domain by applying the independent component analysis (ICA). The basis functions of the air and bone conduction voices extracted by applying ICA are shown first, and then the quality of the signal reconstruction by using several of those basis functions is investigated from the signal compression viewpoint.**

Keywords: ***air conduction voice, bone conduction voice, feature extraction, independent component analysis, signal reconstruction***

1 Introduction

A voice signal has various features both in the time domain and in the frequency domain. The voice signal can be expressed compactly with a small amount of information if those features are grasped precisely and used effectively. Compact expression of voice signal is important in communication from the data compression viewpoint.

Here we consider an air conduction voice and a bone conduction voice. The air conduction voice is recorded by a normal condenser microphone. The bone conduction voice is recorded by a bone conduction microphone, which eliminates surrounding noise and enables communication by voice in extremely noisy environments [1], [2]. It detects the vibration of bones such as jaws and converts its vibration to voice.

The sparse coding and the independent component analysis (ICA) are well known as feature extraction methods using only observed signals. The sparse coding is a method to represent an image signal by using a few basis functions extracted from natural images [3], [4]. This is based on the perceptual system of the mammalian visual cortex. The ICA is closely related to the sparse coding. This is a statistical method to estimate underlying features of the observed signal [5]-[7].

In this paper, the basis functions are first extracted from the air conduction voice and from the bone conduction voice with use of ICA. A voice signal in a short time interval is then reconstructed by using the basis functions extracted, and discussions are given from the point of signal compression.

2 Feature Extraction and Signal Reconstruction of Voice Signal by ICA

2.1 Representation of Voice Signal

A voice signal can be approximated as a stationary signal in a short time interval (tens of milliseconds) though it is regarded as non-stationary in a long time interval [8]. So the voice signal $\boldsymbol{x}$ in a short time interval can be represented as a linear sum of the basis functions $\boldsymbol{a}_i$ as follows [9]:

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

where s_i are the coefficients for each basis function $\boldsymbol{a}_i$, and N is the number of the basis functions. $\boldsymbol{x}$ and $\boldsymbol{a}_i$ are given by:

$$\boldsymbol{x} = (x_1, \ldots, x_M)^T, \tag{2}$$

$$\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iM})^T, \tag{3}$$

where M is a dimension of $\boldsymbol{x}$ sampled in the short time interval, and T is a transpose operation. In the case where $M < N$, it is called the overcomplete representation [10], [11]. Here, for simplicity, we assume that M=N.

2.2 Feature Extraction

(1) can be expressed in the matrix form as follows:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{4}$$

*Manuscript received December 25, 2007. This work was supported by Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B), 19300078.

T. Azetsu is with Yamaguchi Prefectural University, 3-2-1 Sakurabatake, Yamaguchi 753-8502, Japan (phone: +81-83-928-4782; fax: +81-83-928-2251; e-mail: azetsu@yamaguchi-pu.ac.jp).

E. Uchino, R. Kubota, and N. Suetake are with the Graduate School of Science and Engineering, Mathematics and Complex Systems Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan (phone: +81-83-933-5699; fax: +81-83-933-5699; e-mail: uchino@yamaguchi-u.ac.jp).

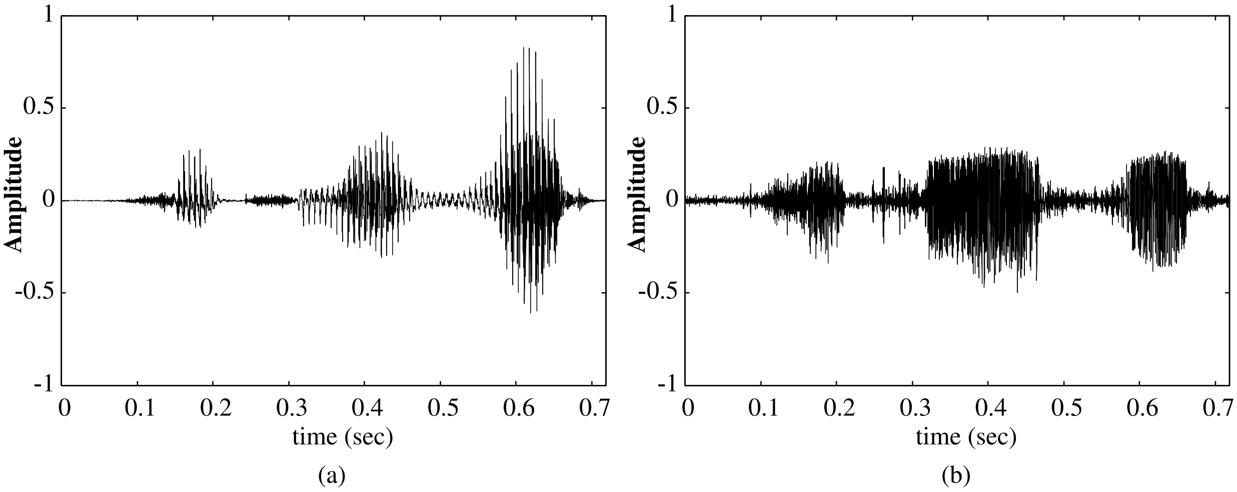

Figure 1: A sample pair of air and bone conduction voices employed for the simulation. These two voices are the Japanese word "hachinohe" spoken by a male speaker. (a) Air conduction voice. (b) Bone conduction voice.

where $\boldsymbol{A}$ and $\boldsymbol{s}$ are given by:

$$\boldsymbol{s} = (s_1, \ldots, s_N)^T, \tag{5}$$

$$\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N). \tag{6}$$

The basis functions $\boldsymbol{a}_i$ and the coefficients s_i are statistically determined from a set of short voice signal $\boldsymbol{x}$ which is cut out from a long voice signal at random. As $\boldsymbol{a}_i$ and s_i are estimated only by using $\boldsymbol{x}$, this is considered to be a blind signal separation (BSS) problem. The ICA is a statistical method to solve the BSS problem by applying a linear transformation to $\boldsymbol{x}$:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}, \tag{7}$$

where $\boldsymbol{y}$ is the separated signal, and $\boldsymbol{W}$ is called an unmixing matrix of size $N \times N$. $\boldsymbol{W}$ is determined by learning so that the components of $\boldsymbol{y}$ become statistically independent of each other.

In the major ICA algorithm, the learning rule of $\boldsymbol{W}$ is derived by applying the natural gradient method as follows [12], [13]:

$$\Delta\boldsymbol{W} = \eta\left(\boldsymbol{I} - \boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T\right)\boldsymbol{W}, \tag{8}$$

where $\Delta\boldsymbol{W}$ is an updated value of $\boldsymbol{W}$, η is a learning rate, and $\boldsymbol{\phi}(\boldsymbol{y})$ is given by:

$$\boldsymbol{\phi}(\boldsymbol{y}) = (\phi(y_1), \ldots, \phi(y_M)). \tag{9}$$

The nonlinear function $\phi(y_j)$ is often described by the following sigmoid type function [14]:

$$\phi(y_j) = -1 + \frac{2}{1 + \exp(-y_j)}. \tag{10}$$

2.3 Signal Reconstruction

When $\boldsymbol{x}$ is sphered (uncorrelated and with unit variances), $\boldsymbol{W}$ can be limited only to orthogonal matrices. $\boldsymbol{W}^T$ is then an estimation of $\boldsymbol{A}$. Also the coefficients $\boldsymbol{s}$=$(s_1, \ldots, s_N)$ are calculated as follows:

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}. \tag{11}$$

The reconstructed voice signal can be obtained by substituting $\boldsymbol{W}^T$ and (11) to (1).

3 Simulation Results

We have performed computer simulations using an air conduction voice and a bone conduction voice. Fig. 1 shows the sample voice signals used for the simulation. These two voices are the Japanese word "hachinohe" spoken by a male speaker recorded by a normal condenser microphone (air conduction voice) and by a bone conduction microphone (bone conduction voice).

It is observed that the bone conduction voice is distorted compared with the air conduction voice. This is because that the bone conduction microphone detects the vibration of bones.

Table 1: Parameters employed for the ICA learning.

Dimension of basis function	64 (8ms)
Number of basis functions	64
Number of $\boldsymbol{x}$	10,000
Learning rate η	0.0001
Nonlinear function $\phi(y_j)$	$-1 + 2/(1 + \exp(-y_j))$
Number of iterations	1,000

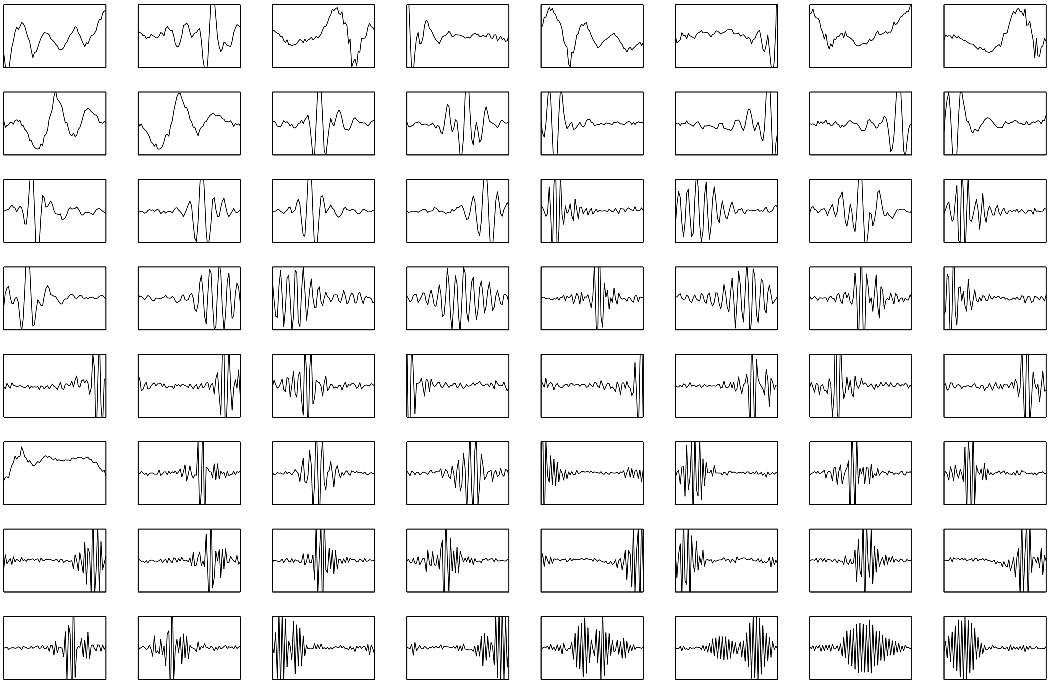

Figure 2: The basis functions of an air conduction voice extracted by ICA.

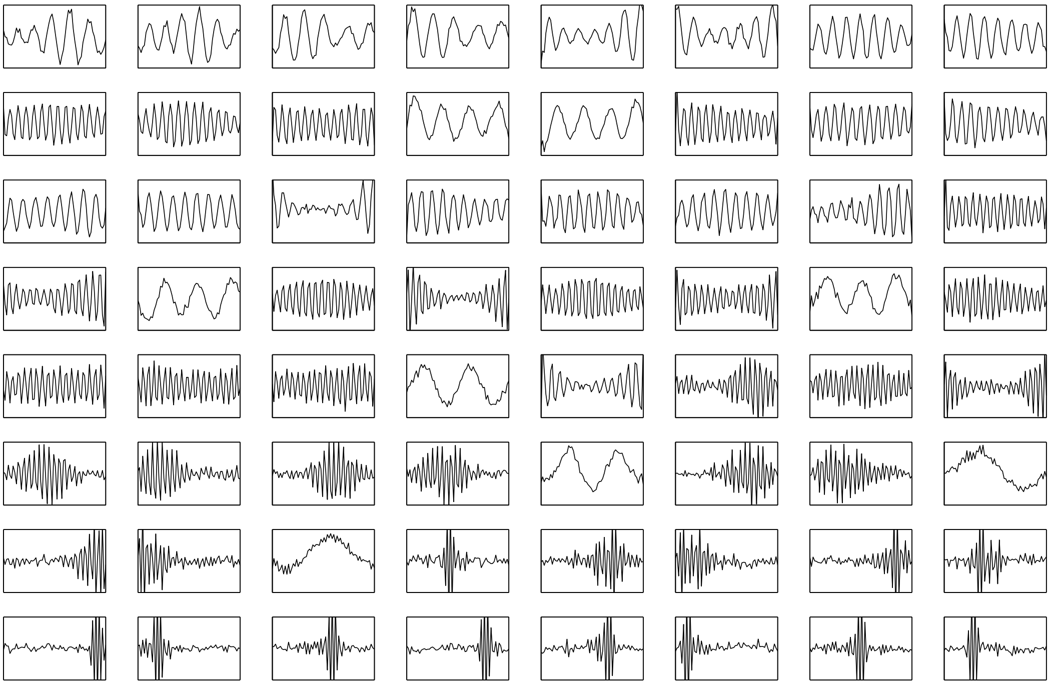

Figure 3: The basis functions of a bone conduction voice extracted by ICA.

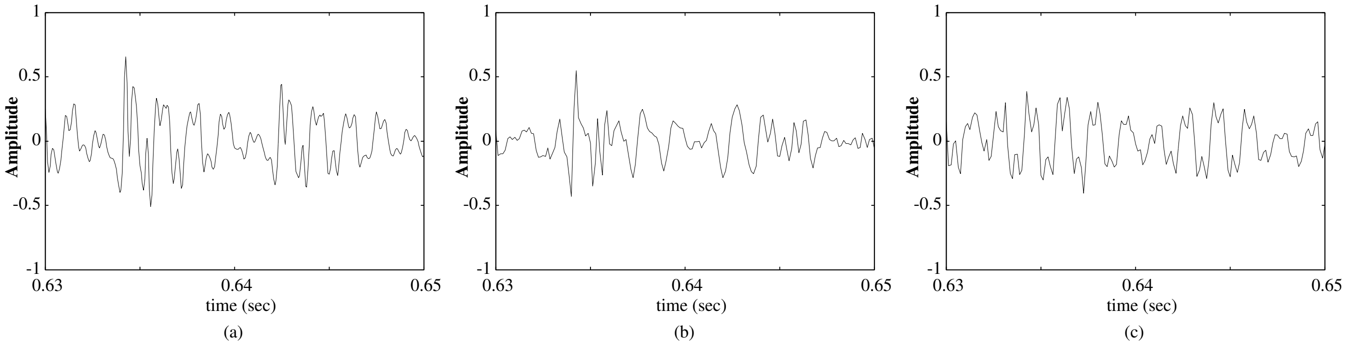

Figure 4: Voice reconstruction results. (a) Part of the Japanese word “hachinohe” of the air conduction voice of Fig. 1 (a). (b) Reconstruction by using the air conduction voice basis functions of Fig. 2. (c) Reconstruction by using the bone conduction voice basis functions of Fig. 3. The number of basis functions used is 6 for both cases.

Table 2: The normalized mean square error (NMSE) of the Japanese word “hachinohe” of the air conduction voice reconstructed by using the basis functions of the air conduction voice and by using those of the bone conduction voice. The basis functions are selected in descending order of contribution, i.e., the absolute value of the corresponding coefficient s_i.

Number of basis functions	NMSE	
	Air conduction basis functions	Bone conduction basis functions
1	0.9046	0.8542
3	0.7662	0.7039
6	0.6211	0.5575
9	0.5129	0.4625
12	0.4272	0.3940

In the simulations, 100 Japanese words were used which were recorded both by a condenser microphone (air conduction voice) and by a bone conduction microphone (bone conduction voice) at the same time. Each word is about 1 sec. long. 99 Japanese words among those are used to extract the basis functions by ICA. The remaining one word is used for a signal reconstruction by using those basis functions extracted. Table 1 shows the parameters employed for the ICA learning.

Figs. 2 and 3 show the basis functions extracted from the air and from the bone conduction voices, respectively. From those figures, it is observed that the basis functions extracted from the bone conduction voice have more periodic signals compared with those extracted from the air conduction voice.

The accuracy of signal reconstruction by using the basis functions of the air or of the bone conduction voice is quantitatively evaluated by the normalized mean square error (NMSE) defined by:

$$\mathrm{NMSE} = \frac{\sum_t (x(t) - x^*(t))^2}{\sum_t x(t)^2}, \qquad (12)$$

where $x(t)$ is an original signal, $x^*(t)$ is its reconstructed signal, and t indicates a discrete time.

Table 3: The NMSE of the Japanese word “hachinohe” of the bone conduction voice reconstructed by using the basis functions of the air conduction voice and by using those of the bone conduction voice. The basis functions are selected in descending order of contribution, i.e., the absolute value of the corresponding coefficient s_i.

Number of basis functions	NMSE	
	Air conduction basis functions	Bone conduction basis functions
1	0.8888	0.8134
3	0.7447	0.6352
6	0.5954	0.4912
9	0.4929	0.3987
12	0.4147	0.3299

Table 2 shows the NMSE of the Japanese word “hachinohe” of the air conduction voice reconstructed by using the basis functions of the air conduction voice of Fig. 2 and by using those of the bone conduction voice of Fig. 3. The NMSEs with changing the number of the basis functions are shown in the table.

Worth nothing is that, even the target signal to be reconstructed is an air conduction voice, the set of the basis functions of the bone conduction voice gives better reconstruction results than that of the air conduction voice.

Table 3 shows the reconstruction results of the bone conduction voice as opposed to the results of the air conduction voice of Table 2.

Fig. 4 shows the Japanese word “hachinohe” of the air conduction voice reconstructed by using the basis functions of the air and of the bone conduction voices, respectively. The number of basis functions used is 6.

4 Conclusions

ICA has been applied to the extraction of the basis functions for expressing the air and the bone conduction voices. We have reconstructed Japanese words by using those basis functions extracted.

The simulation results have given an interesting fact that the basis functions extracted from the bone conduction voice give good reconstruction results even for the reconstruction of the air conduction voice. This means that the air conduction voice (recorded by the normal condenser microphone) can be effectively represented by the bone conduction basis functions. It is expected from the point of data compression that this leads to the compact coding of the conversational voice for telecommunication.

References

[1] Uchino, E., Yano, K., Azetsu, T., “Twin units self-organizing map for voice conversion,” *Electronics Letters*, 39(24), pp. 1767-1769, 2003.

[2] Uchino, E., Yano, K., Azetsu, T., “A self-organizing map with twin units capable of describing a nonlinear input-output relation applied to speech code vector mapping,” *Information Sciences*, 177, pp. 4634-4644, 2007.

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

[4] Olshausen, B. A., Field, D. J., “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” *Vision Research*, 37, pp. 3311-3325, 1997.

[5] Bell, A. J., Sejnowski, T. J., “The ‘independent components’ of natural scenes are edge filters,” *Vision Research*, 37, pp. 3327-3338, 1997.

[6] Hyvärinen, A., Hoyer, P. O., “Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces,” *Neural Computation*, 12(7), pp. 1705-1720, 2000.

[7] Hyvärinen, A., Karhunen, J., Oja, E., *Independent Component Analysis*, John Wiley, 2001.

[8] O’shaughnessy, D., *Speech Communications*, IEEE Press, 2000.

[9] Kotani, M., Shirata, Y., Maekawa, S., Ozawa, S., Akazawa, K., “Representations of speech by sparse coding algorithm,” *IEEJ Trans. on Electronics, Information and Systems*, 120-C(12), pp. 1996-2002, 2000 (in Japanese).

[10] Lee, T. -W., Lewicki, M. S., Girolami, M., Sejnowski, T. J., “Blind source separation of more sources than mixtures using overcomplete representations,” *IEEE Signal Processing Letter*, 6(4), pp. 87-90, 1999.

[11] Lewicki, M. S., Sejnowski, T. J., “Learning overcomplete representations,” *Neural Computation*, 12(2), pp. 337-365, 2000.

[12] Amari, S., Cichocki, A., Yang, H. H., *A new learning algorithm for blind signal separation*, in: Advances in Neural Information Processing Systems 8, MIT Press, pp. 757-763, 1996.

[13] Amari, S., “Natural gradient works efficiently in learning,” *Neural Computation*, 10, pp. 251-276, 1998.

[14] Bell, A. J., Sejnowski, T. J., “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, 7, pp. 1129-1159, 1995.