CONCURRENT SPEECH SIGNAL SEPARATION BASED ON FREQUENCY DOMAIN BINAURAL MODEL

Yoshifumi Chisaki, Takashi Nakanishi, Hidetoshi Nakashima and Tsuyoshi Usagawa*

Dept. of Computer Science, Kumamoto Univ. 2-39-1 Kurokami, Kumamoto, 860-8555 Japan
email: *chisaki@cs.kumamoto-u.ac.jp*
* Kumamoto National College of Technology, Kumamoto, 861-1102 Japan

ABSTRACT

We can communicate with others in a noisy environment. This phenomenon is known as a "Cocktail Party Effect," and is one of the most important binaural functions. In our previous study, frequency domain binaural model (FDBM) based on interaural phase difference (IPD) and interaural level difference (ILD) was proposed, and its performance as the front-end of the speech recognition system was confirmed. However, the model estimates direction of arrival (DOA) of the sound sources only in azimuth direction. This paper addresses the extended model which estimates DOA in both azimuth and elevation, simultaneously.

The proposed model is evaluated not only as a speech enhancer but also as a front-end of the speech recognition system. According to the evaluation as a speech enhancer, the envelope of the segregated signal is recovered and quite similar to the one of the target signal. On the other hand, more than 90% recognition rates are obtained in speech recognition task, when the azimuth of reception of the target signal and noise differs by 10˚.

1. INTRODUCTION

There are a lot of issues for concurrent noise reduction including speech signal since several applications require these techniques, such as speech recognition system. The microphone array system is one of them. However, there are some restrictions such as number of sound sources to be separated and computational load.

This work was partly supported by Grant-in-Aid for Scientic Research (C) No.14550422 and by the Research Institute of Electrical Communication, Tohoku University.

On the other hand, there are other approaches for signal segregation. The cocktail party processor [1] based on the time domain binaural model [2] is one of them. This model utilizes an interaural cross-correlation of each critical band to extract signal components from specific direction and segregates the concurrent speeches. However, this model requires huge computational load because of time domain processing and it is very hard to make it a real time application. In an attempt to realize handy application, the frequency domain binaural model [3] (FDBM) based on interaural phase difference (IPD) and interaural level difference (ILD) was proposed by the authors. Although this model can segregate the sound source in a specific direction, it does not take into account elevation as the DOA of the sound source.

This paper addresses the FDBM which performs DOA estimation in azimuth and elevation. The performance of the proposed FDBM is evaluated with both speech segregation task and speech recognition task.

2. METHOD

In the proposed method, signals are observed by using dummy-head. Let's assume that the target signal denotes $s_m(n)$ ($m = 0$), and the interference signals also denote $s_m(n)$ ($m = 1, 2, 3....$), the observed signals, $l(n)$ and $r(n)$, can represent as

$$\begin{aligned} l(n) &= l_0(n) + l_1(n) + l_2(n) + ... \\ &= \sum_m s_m(n) * h_{l,m}(n) \\ r(n) &= r_0(n) + r_1(n) + r_2(n) + ... \\ &= \sum_m s_m(n) * h_{r,m}(n), \end{aligned}$$

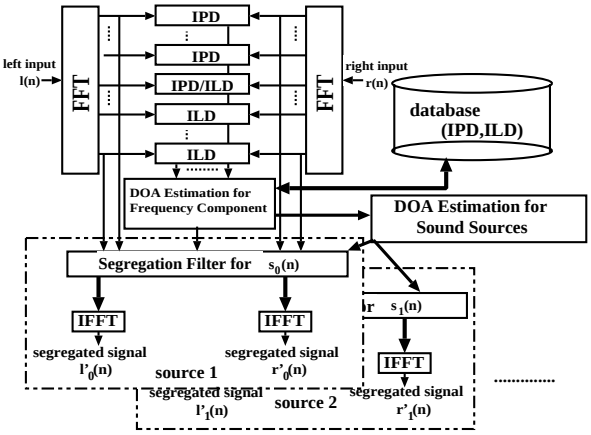

Fig. 1. Block diagram of the FDBM (Frequency Domain Binaural Model).

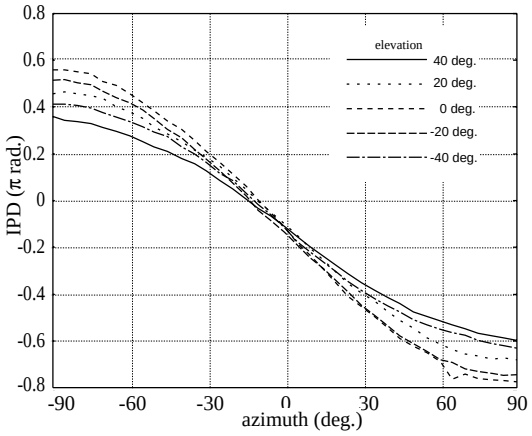

Fig. 2. IPD-DOA map of the HRTFs relative to the azimuth and elevation for 150Hz. Note that 0˚ of the azimuth and elevation indicates in front of the dummy-head.

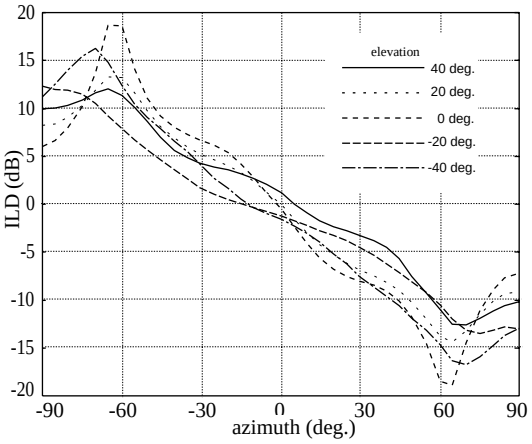

Fig. 3. ILD-DOA map of the HRTFs relative to the azimuth and elevation for 2kHz. 0˚ indicates in front of the dummy-head.

where $h(n)$ represents head-related transfer function (HRTF) from appropriate direction, and * denotes convolution operator. Figure 1 shows the block diagram of FDBM. The main-block of proposed model consists of several sub-blocks. The role of each block is as follow.

2.1. FFT Analysis

The both input signals, $l(n)$ and $r(n)$, observed using the microphones attached in a ear canal of the dummy-head are transformed to obtain spectra, $L(k)$ and $R(k)$, from the time domain into the frequency domain by FFT of 512 taps.

2.2. DOA Estimation by IPD

The IPD carries the DOA information of each frequency component, mainly lower frequency bands. It can be obtained through a cross spectrum, $C_{lr}(k)$, defined as follows,

$$C_{lr}(k) = L(k)R(k)^*, \tag{1}$$

where * denotes complex conjugate. And the IPD, $\theta_{lr}(k)$, in each frequency is obtained using the cross spectrum, $C_{lr}(k)$, as follow,

$$\theta_{lr}(k) = \tan^{-1}\left\{\frac{\mathrm{Im}[C_{lr}(k)]}{\mathrm{Re}[C_{lr}(k)]}\right\}. \tag{2}$$

DOA information $D(k,\phi,\psi)$ of each frequency component is determined by comparing with the IPD-DOA map, $\theta_{map}(k,\phi,\psi)$, obtained *a priori* using HRTFs. Note that ϕ and ψ represent azimuth and elevation, respectively. Figure 2 shows the IPD-DOA map for 150Hz. The vertical and horizontal axis represent IPD and sound source azimuth direction, respectively, while the 0˚ indicates in front of dummy-head. The difference between IPD and IPD-DOA map is obtained on each azimuth and elevation as

$$\Delta\theta(k,\phi,\psi) = |\theta_{lr}(k) - \theta_{map}(k,\phi,\psi)|, \tag{3}$$

and DOA information is defined as

$$D_{IPD}(k,\phi,\psi) = e^{-\alpha_1(k)\cdot\Delta\theta(k,\phi,\psi)}, \tag{4}$$

where $\alpha_1(k)$ is a weighting function depending on the frequency.

2.3. DOA Estimation by ILD

The ILD also carries DOA information for each frequency component as well as IPD. However, in low frequency band, the ILD is quite small so that the sound in the band is well diffracted by head. The sound in high frequency band, on the other hand, cannot be diffracted, and ILD becomes large. The ILD, $\xi_{lr}(k)$, for each frequency component is obtained as

$$\xi_{lr}(k) = 20 \log \left| \frac{C_{lr}(k)}{C_{ll}(k)} \right|, \qquad (5)$$

where $C_{ll}(k)$ represents power spectrum of $L(k)$. $\xi_{lr}(k)$ is utilized to determine the DOA by comparing it with the ILD-DOA map $\xi_{map}(k, \phi, \psi)$ based on following definitions.

$$\Delta\xi(k, \phi, \psi) = |\xi_{lr}(k) - \xi_{map}(k, \phi, \psi)|, \qquad (6)$$

and the DOA information based on ILD is defined as

$$D_{ILD}(k, \phi, \psi) = e^{-\alpha_2(k) \cdot \Delta\xi(k,\phi,\psi)}, \qquad (7)$$

where $\alpha_2(k)$ is also a weighting function. Figure 3 shows the ILD-DOA map for 2kHz. The vertical and horizontal axis represent ILD and sound source azimuth direction, respectively, while the 0˚ indicates in front of dummy-head.

2.4. DOA Estimation for Sound Source

The obtained DOA information based on IPD and ILD for each frequency components are combined as

$$\begin{aligned} D(k, \phi, \psi) &= (1 - \beta(k)) \cdot D_{ILD}(k, \phi, \psi) \\ &\quad + \beta(k) \cdot D_{IPD}(k, \phi, \psi), \end{aligned} \qquad (8)$$

where $\beta(k)$ represents forgetting factor depending on frequency. For example, $\beta(k)$=0 for low frequency(0 to 1kHz), $\beta(k)$=0 to 1 (varies gradually) for 1 to 2kHz, and $\beta(k) = 1$ for more than 2kHz. The DOA of the "m"th sound source is estimated by using following equation.

$$\begin{aligned} D_{OA}(m) &= \{\phi_m, \psi_m| \\ &\quad max(\sum_k E(k) \cdot D(k, \phi, \psi))\}, \end{aligned} \qquad (9)$$

where $E(k)$ represents energy-dependent weighting factor on the frequency k.

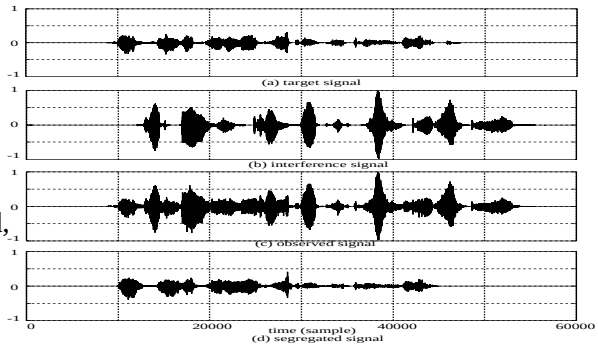

Fig. 4. Input and output waveforms as results of the speech segregation task.

2.5. Signal Segregation

The segregation filter $H_m(k)$ to segregate "m"th sound source is defined using estimated DOA information, ϕ_m and ψ_m, as

$$H(k) = D(k, \phi_m, \psi_m), \qquad (10)$$

and the segregated signal $l'_m(n)$ and $r'_m(n)$ is obtained as

$$l'_m(n) = IFFT[L(k)H(k)], \qquad (11)$$

$$r'_m(n) = IFFT[R(k)H(k)]. \qquad (12)$$

Note that the binaural information in the segregated signals are preserved.

3. COMPUTER SIMULATIONS

In this section, the results of the computer simulation are shown. The simulations are (1) Signal segregation task, and (2) Speech recognition task. In each simulations, the sampling frequency of the signal is set to 16kHz, and the HRTFs of KAMER dummy-head microphone, which is provided by MIT Media Lab[4], are utilized for IPD-DOA and ILD-DOA maps.

3.1. Signal Segregation Task

Signal segregation is performed by using the proposed method. Situation of the simulation is that the target signal comes from ($\phi_0 = -30^\circ$, $\psi_0 = -20^\circ$), and the interference signal comes from ($\phi_1 = 30^\circ$, $\psi_1 = 20^\circ$). The SNR between these signals is set to 0dB. Figure 4 shows the result of this simulation. The waveforms

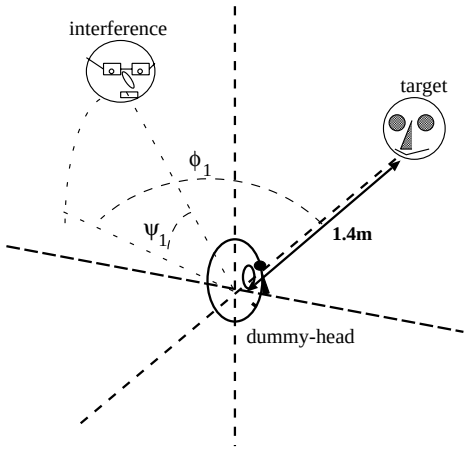

Fig. 5. Relative location between dummy-head and sound sources in speech recognition task.

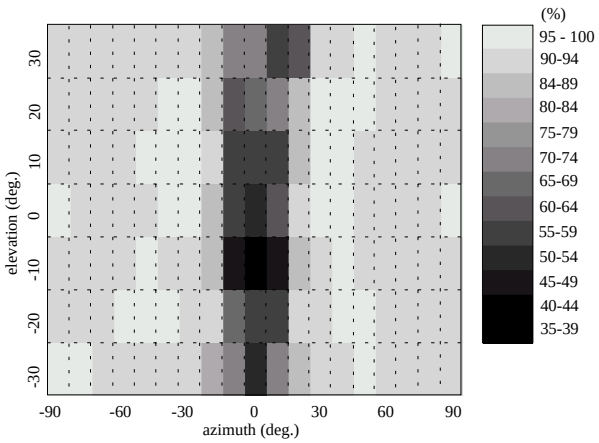

Fig. 6. Result of the speech recognition task.

show (a) target signal, (b) interference signal, (c) observed signal, and (d) segregated target signal. In this figure, only the Rch signals are shown. Thus the power ratio between (a) and (b) is less than 0dB. However, the envelope of (d) segregated signal is quite similar to (a) target signal. From this figure, signals can be segregated by using the proposed method even if the elevation varies.

3.2. Speech Recognition Task

To confirm the performance of the proposed method as a front-end of the speech recognition system, speech recognition task is performed. Relative location between dummy-head and sound sources is shown in Fig. 5. The target speech signal comes from in front of the dummy-head ($\phi_0 = 0^\circ$, $\psi_0 = 0^\circ$), while DOA of the interference speech signal(ϕ_1, ψ_1) varies from -90° to $+90^\circ$ for azimuth and -30° to $+30^\circ$ for elevation in 10° step. Female voice is employed as the target signal and male voice is as the interference signal. SNR between these signals is set to 0dB.

Figure 6 shows the recognition rate according to the DOA of the interference. Horizontal and vertical axis indicate azimuth and elevation of the interference, respectively. When the interference closes to the target, the rate is down to 40% approximately. However, the angle between the interference and target is more than 20° in azimuth, the rate is more than 90% in any case. This tendency means that FDBM segregates the sound sources when the azimuth of the target and interference is apart.

4. SUMMARY

In this paper, FDBM, taking into account elevation for the DOA, is proposed because DOAs of the sound sources are not always in same elevation in real situations. From results of the computer simulation, FDBM not only works as a speech enhancer but also as a front-end of the speech recognition or sound segregation system. Particularly as a front-end of the speech recognition system, the proposed method shows more than 90% recognition rates even if the elevation of the sound source is varied when the azimuth of reception of the target signal and noise differs by 10°.

5. ACKNOWLEDGEMENT

This work was partly supported by Grant-in-Aid for Scientic Research (C) No.14550422 and by the Research Institute of Electrical Communication, Tohoku University.

6. REFERENCES

[1] M. Bodden, *Acta Acustica*, Vol.1, pp.43–55 (1993)

[2] W. Lindemann, *J. Acoust. Soc. Am.*, Vol.80, pp.1608–1622 (1986)

[3] H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, *Acoust. Sci. & Tech.* Vol.24, pp.172–178 (2003)

[4] http://sound.media.mit.edu/KEMAR.html