Concurrent Speech Segregation based on DOA Information using Frequency Domain Binaural Model
– An application for hearing aid –

Tsuyoshi Usagawa, Rika Matsuo, Takashi Nakanishi, Hidetoshi Nakashima*, Yoshifumi Chisaki

Department of Computer Science, Kumamoto University, Japan
* Kumamoto National College of Technology, Japan
{tuie, chisaki}@cs.kumamoto-u.ac.jp, *nakashi@ec.knct.ac.jp

Abstract
A Frequency domain binaural model (FDBM) can provide an azimuth and an elevation of each signal source as a direction of arrival (DOA) from the binaural input of dummy head. Based on the DOA information, this model can segregate a few concurrent speeches as a human being. The performance of segregation of FDBM is measured both as an front-end of automatic speech recognition system and hearing-aid as well as an speech enhancer for transmission and recording purpose. Improvement obtained by FDBM front-end is more than 15dB when DOAs of concurrent speeches are differed more than 30˚ under SNR=0dB.

1. Introduction
The microphone array system is one of the typical methods for signal segregation or signal enhancement when multiple signal sources are spread ed in the environment. Although it is an attractive method for various applications, it has serious restriction on a number of elements and computational load.

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 an example. This model utilizes an interaural cross-correlation on 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. 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 not only in azimuth but also in elevation. The performance of the proposed FDBM is evaluated with both speech segregation task and speech recognition task. And the preliminary implementation of binaural hearing aid is discussed by the directivity pattern.

2. METHOD
Let’s assume that the target signals are denoted as \( s_m(n) \) \( (m = 1, 2, 3, \ldots ) \). The observed signals by left and right ear position, \( l(n) \) and \( r(n) \), can represent as
\[
l(n) = l_1(n) + l_2(n) + l_3(n) + \ldots = \sum_{m} s_m(n) * h_{l,m}(n) \]
\[
r(n) = r_1(n) + r_2(n) + r_3(n) + \ldots = \sum_{m} s_m(n) * h_{r,m}(n),
\]
where \( h_{l,m}(n) \) and \( h_{r,m}(n) \) represent head-related transfer functions (HRTFs) from the direction of \( m \)-th source to “left” and “right” ear positions, and \(*\) means convolution. Figure 1 shows the block diagram of FDBM. The main-block of proposed model consists of several sub-blocks described as follows.

2.1. FFT Analysis
The both input signals, \( l(n) \) and \( r(n) \), observed by microphones attached to the dummy-head are transformed into spectra, \( L(k) \) and \( R(k) \), by means of FFT.

2.2. DOA Estimation by IPD
For lower frequency bands, IPD mainly carries the DOA information. It can be obtained through a cross spectrum, \( C_{lr}(k) \), defined as follows,
\[
C_{lr}(k) = L(k)R(k)^*,
\]
where \(*\) denotes complex conjugate. And the IPD, \( \theta_{lr}(k) \), in each frequency component \( k \) is obtained using the cross spectrum, \( C_{lr}(k) \), as follow,
\[
\theta_{lr}(k) = \tan^{-1}\left\{ \frac{\text{Im}[C_{lr}(k)]}{\text{Re}[C_{lr}(k)]} \right\}.
\]
IPD-based DOA information $D_{IPD}(k, \phi, \psi)$ of each frequency component is determined by comparing $\theta_{lr}(k)$ with the IPD-DOA map, $\theta_{map}(k, \phi, \psi)$, obtained a priori using HRTFs. Note that $\phi$ and $\psi$ represent azimuth and elevation, respectively. Figure 2 shows the IPD-DOA map at 150Hz for various azimuth and elevation. The vertical and horizontal axes represent IPD and sound source azimuth direction, respectively, while the $0^\circ$ 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
\[
D_{IPD}(k, \phi, \psi) = e^{-\alpha_1(k) \Delta \theta(k, \phi, \psi)}, \tag{4}
\]
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 bands, the ILD is quite small because the low frequency components are well diffracted by head. In higher frequency range, on the other hand, ILD becomes large. The ILD, $\xi_{lr}(k)$, for each frequency component is obtained as
\[
\xi_{lr}(k) = 20 \log \left| \frac{C_{lr}(k)}{C_{ll}(k)} \right|, \tag{5}
\]
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.
\[
\Delta \xi(k, \phi, \psi) = |\xi_{lr}(k) - \xi_{map}(k, \phi, \psi)|, \tag{6}
\]
and the DOA information based on ILD is defined as
\[
D_{ILD}(k, \phi, \psi) = e^{-\alpha_2(k) \Delta \xi(k, \phi, \psi)}, \tag{7}
\]
where $\alpha_2(k)$ is also a weighting function. Figure 3 shows the ILD-DOA map for 2kHz. The vertical and horizontal axes represent ILD and sound source azimuth direction, respectively, while the $0^\circ$ 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
\[
D(k, \phi, \psi) = (1 - \beta(k)) \cdot D_{ILD}(k, \phi, \psi) + \beta(k) \cdot D_{IPD}(k, \phi, \psi), \tag{8}
\]
where $\beta(k)$ represents forgetting factor depending on frequency. For example, $\beta(k)=0$ for less than 1kHz, $\beta(k)$ varies gradually from 0 to 1 according to the frequency from 1kHz to 2kHz, and $\beta(k) = 1$ for more than 2kHz.
The DOA of the $m$-th sound source is estimated by using following equation.

$$D_{OA}(m) = \{\phi_m, \psi_m\}$$

$$\max\left(\sum_k E(k) \cdot D(k, \phi, \psi)\right),$$

where $E(k)$ represents energy-dependent weighting factor on the frequency $k$. It means that DOA for $m$-th source is determined as the best matched combination of $\phi$ and $\psi$ over several frequency bands.

### 2.5. Signal Segregation

The segregation filter $H_m(k)$ to segregate $m$-th sound source is defined using estimated DOA information, $\phi_m$ and $\psi_m$, as

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

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

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

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

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

### 3. Computer Simulations

In this section, the results of two computer simulations, (1) signal segregation and (2) speech recognition tasks, are shown. 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[5], are utilized for IPD-DOA and ILD-DOA maps.

#### 3.1. Signal Segregation Task

Signal segregation is performed by using the proposed method. Condition of the simulation is that the target signal comes from ($-30^\circ$, $-20^\circ$), and the interference signal comes from ($30^\circ$, $20^\circ$). The SNR between these signals is set to 0dB. Figure 4 shows the result of this simulation. The waveforms show (a) target signal, (b) interference signal, (c) observed signal, and (d) segregated target signal. Although target (a) is weaker than interference (b), the envelope of segregated signal (d) is similar to (a) target.

#### 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 ($0^\circ$, $0^\circ$), while DOA of the interference speech signal varies from $-90^\circ$ to $+90^\circ$ for azimuth and $-30^\circ$ to $+30^\circ$ for elevation in $10^\circ$ step. Female voice is employed as the target signal and male voice is as the interference signal. SNR between these signals is 0dB.

Figure 6 shows the recognition rate according to the DOA of the interference. Horizontal and vertical axes indicate azimuth and elevation of the interference, respectively. The rate becomes about 40% when DOA of interference closes to the target. However, when the angle between the interference and target is more than $20^\circ$ 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. Application for Hearing Aid System

Figure 7 shows an experimental setup for a binaural hearing aid system based on FDBM. Microphones are attached on the outer cover of headphone. In order to show the directivity characteristics of the binaural system, the howling canceler is not implemented and the headphone is not activated. HRTF database used in FDBM for this
Figure 6: Result of the speech recognition task.

Figure 7: Experimental setup for headphone type binaural hearing aid experiment is obtained by the set of user’s own HRTFs.

Two kind of speech signals, female and male ones, are used for measuring directivity pattern. Speech signal is reproduced each of 10° step form -90° to +90°. The sampling frequency and resolution for FDBM are set to 16kHz and 16bit, respectively.

The directivity pattern of this hearing aid system is shown in Fig. 8. In this figure, the setting of spatial filtering for azimuth direction is shown as a thin broken line. The obtained directivity patterns are not symmetric due to the reflection at the experimental environment. Although the directivity patterns are less sharp than the setting, the directivity pattern obtained by FDBM shows effectiveness in order to build binaural hearing aid. Because settings of spatial filter can be flexibly changed, a user of hearing aid can adaptively control the directivity pattern according to the situation.

5. Conclusions

In this paper, FDBM which can segregate concurrent speeches based not only on azimuth but also on elevation of DOA is proposed. From results of the computer simulation, FDBM can work not only as a speech enhancer and a front-end of the speech recognition but also as a binaural front end for hearing aid. 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°. Also as the front of hearing aid, it can enchance the speech signal about 10dB for specified DOA of the user’s coordinate.

6. Acknowledgments

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

7. References


