Binaural Speech Segregation System on Single Board Computer

Tsuyoshi USAGAWA
Kumamoto University, 2-29-1 Kurokami, Chuo, Kumamoto, 860-8555, Japan,
tuie@cs.kumamoto-u.ac.jp

IRWANSYAH
Kumamoto University, 2-29-1 Kurokami, Chuo, Kumamoto, 860-8555, Japan,
irwan@hicc.cs.kumamoto-u.ac.jp, iru.one.syah@gmail.com

Abstract

A pocket-size binaural speech segregation system has been developed and assembled with available consumer hardware. It can enhance a target speech in a certain direction while attenuating interfering sounds from other directions. This system is based on the frequency domain binaural model (FDBM) and it segregates multiple speeches based on the directivity. This real-time system is implemented on a low-cost single board computer which might be used as a hearing assistance device. Performance of the system is evaluated in a normal laboratory room as well as anechoic chamber. Even in a room with reverberation, the system works well and show the almost same performance obtained in anechoic chamber.

Keywords: Frequency Domain Binaural Model, Single Board Computer, Hearing Assistance, Open-Source Software

1. Introduction

Very rapid aging of society in Japan, hearing assistant systems and hearing aids attract the attention of senior members as well as ones who have hearing disables. There are several very sophisticated hearing aids, namely digital ones, on a market, however, comparing the senior glasses, the market price has a significant differences between hearing aids and senior glasses. Also monaural hearing assistance does not always work well as monocular glass does not, thus it is recommended to use a pair of hearing aids for binaural assistance, which costs almost the double. Beside other reasons, this price gap makes a hearing aid difficult to be popular.

In this paper, the development of an "open-source" hearing assistance device with consumer hardware in affordable price range. The proposed hearing device is intended to assist a listener to focus her/his attention on a speaker that she/he is talking with. This binaural system can segregate and enhance sounds which come from specified direction using binaural cues by a frequency domain binaural model (FDBM) proposed by Nakashima *et al*. in 2003. This model uses the binaural cues, namely interaural level difference (ILD) and interaural phase difference (IPD) to estimate directions of sources in multiple sound source condition and segregates sounds based on estimated directions. Also, since FDBM works in the frequency domain, it has a relatively low computational

complexity compared to time domain based models. Using this advantage, FDBM is implemented on a cost-effective single board computers (SBCs) and it is controlled by Android-based smart phone by an application named "OpenFDBM" which allows a user to control the hearing assistance device via graphical interface. The performance of implemented FDBM as a binaural hearing assistance system is measured in an anechoic chamber as well as ordinary laboratory room with reverberation. Source code of FDBM written in Python which has rich audio signal processing libraries, run as an real-time application on SBC. Source codes including instructions to build the hearing device are freely available on Github and demonstration video is also available online.

2. Frequency Domain Binaural Model

This section shortly describes the overall overview of the original frequency domain binaural model [1]. However, here the original algorithm is slightly modified to use zero padding of fast Fourier transform (FFT) in order to maintain a high-frequency resolution when using a shorter frame length. Figure 1 shows the overall FDBM-based speech segregation scheme. Let $x_R(i)$ and $x_L(i)$ be the observed signals received at left and right microphones, defined here as

$$x_R(i) = s(i) * h_{r,\phi_s}(i) + v(i) * h_{r,\phi_v}(i) = s_R(i) + v_R(i) \tag{1}$$

$$x_L(i) = s(i) * h_{l,\phi_s}(i) + v(i) * h_{l,\phi_v}(i) = s_L(i) + v_L(i) \tag{2}$$

where *s*(*i*) and *v*(*i*) are the target and interfering speakers, respectively, and this notation "$*$" indicates the convolution operator. When someone is talking in front of us, the speech signal received by left and right ears would be different, depending on the direction of the sound source. These acoustical differences perceived by left and right ears can be summarized by head related transfer functions (HRTFs). $h_{l,\phi}(i)$ and $h_{r,\phi}(i)$ are then defined as the left and right HRTFs as a function of f azimuth. We divide the FDBM into six stages: Framing Stage, ILD-IPD Calculation, Database Comparison, Weight Combination, Segregation Filter and Separation Stage.

Stage 1. Framing Stage:

For conversion from time domain signals into frequency domain one, a short-time Fourier Transformation (STFT) was used in the original FDBM [2], however a perfect reconstruction method with a half-cycle sine window is introduced [3]. STFTs of observed signals are obtained by FFT for each windowed block:

$$X_R(\lambda, k) = FFT\{x_R^{zp}(i)\};\ X_L(\lambda, k) = FFT\{x_L^{zp}(i)\} \tag{3}$$

where λ is the frame index and k is the frequency index. The zero-padded windowed block, $x_{\{\cdot\}}^{zp}(i)$ is defined as follows

$$x_{\{\cdot\}}^{zp}(i) = \begin{cases} x_{\{\cdot\}}(i) \cdot w(i - \lambda\frac{N}{2}), & 0 \le i \le N-1 \\ 0, & N \le i \le M-1 \end{cases} \tag{4}$$

where N is the window length we want to shorten, and M is the frame length. As discussed in [4], the quality of the segregated sounds depends on the frequency resolution of FFT, which is required at least a 32-ms frame length. For that reason, M is set to 32 ms.

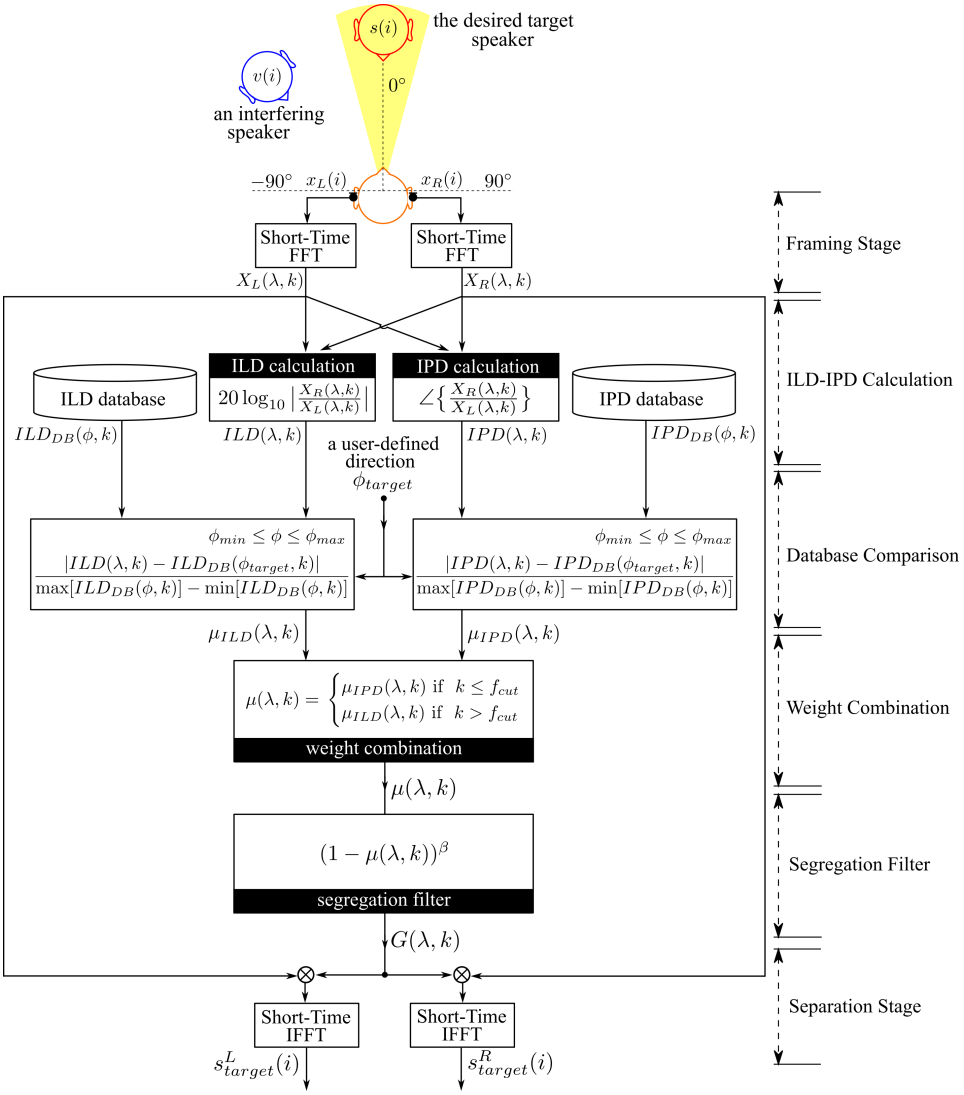

Figure 1. Overall structure of the FDBM-based speech segregation scheme

Stage 2. ILD-IPD Calculation:

It is well known that humans are able to localize a sound source in space by exploring the differences in time and level between the sound reaching the left and right ears. These relative differences are known IPD and ILD, respectively. If HRTFs are available in advance, a target speech signal in a particular direction is possible to be extracted. Using ILD and IPD information, the FDBM is developed to segregate the desired target signal. The ratio of the STFTs of the observed signals, known as the interaural spectrogram $X_{IS}(\lambda,k)$, can be obtained using Eq.(6) and ILD and IPD can be calculated as follows;

$$X_{IS}(\lambda,k)=\frac{X_R(\lambda,k)}{X_L(\lambda,k)} \tag{5}$$

$$ILD(\lambda,k)=20\log_{10}|X_{IS}(\lambda,k)|;\ IPD(\lambda,k)=\angle X_{IS}(\lambda,k) \tag{6}$$

where $\angle$ denotes the phase angle.

Stage 3. Database Comparison:

To extract a target speech signal in a particular direction ϕ_{target}, HRTFs are required by the FDBM. Fortunately, MIT Media Lab has provided HRTFs of KEMAR Dummy-Head microphone [5]. Let $H_{DB,R}(\phi,k)$ and $\mathrm{H}_{DB,L}(\phi,k)$ are right and left HRTF database as a function of azimuth in the frequency domain. Similar to the previous equations, ILD and IPD database can be formulated as

$$ILD_{DB}(\phi,k)=20\log_{10}\left|\frac{H_{DB,R}(\phi,k)}{H_{DB,L}(\phi,k)}\right|;\ IPD_{DB}(\phi,k)=\angle\left\{\frac{H_{DB,R}(\phi,k)}{H_{DB,L}(\phi,k)}\right\} \tag{7}$$

HRTF database provided by MIT covers the full azimuth from 0 to 360 with 5-degree resolution. By comparing ILD and IPD of the observed signals with the ILD and IPD database, weight factors used for extracting the desired target signal in a specific direction target can be obtained. The calculation of weight factors can be defined as

$$\mu_{ILD}(\lambda,k)=\frac{|ILD(\lambda,k)-ILD_{DB}(\phi_{target},k)|}{\max[ILD_{DB}(\phi,k)]-\min[ILD_{DB}(\phi,k)]} \tag{8}$$

$$\mu_{IPD}(\lambda,k)=\frac{|IPD(\lambda,k)-IPD_{DB}(\phi_{target},k)|}{\max[IPD_{DB}(\phi,k)]-\min[IPD_{DB}(\phi,k)]} \tag{9}$$

$\max[f(\phi,k)]$ indicates the maximum value of f at a particular frequency index k when ϕ is varying from ϕ_{min} to ϕ_{max}, depending on the available HRTF database.

Stage 4. Weight Combination:

The weight factors, $\mu_{ILD}(\lambda,k)$ and $\mu_{IPD}(\lambda,k)$, from the previous stage could not work optimally in all frequency region. $\mu_{IPD}(\lambda,k)$ works well only in the low-frequency region. On the other hand, $\mu_{ILD}(\lambda,k)$ is good for the high-frequency region

[1]. Therefore, both need to be combined. The combination of both weight factors is defined as follows:

$$\mu(\lambda,k)=\begin{cases}\mu_{IPD}(\lambda,k) & \text{if} \quad k \le f_{cut}\\ \mu_{ILD}(\lambda,k) & \text{if} \quad k > f_{cut}\end{cases} \tag{10}$$

where f_{cut} is assumed to be the cutoff frequency which we consider to be 1250 Hz. To avoid amplification, the upper limit of the weight factor is set to one.

Stage 5. Segregation Filter:

The segregation filter is then defined as

$$G(\lambda,k)=\left[1-\mu(\lambda,k)\right]^{\beta} \tag{11}$$

where β is the gain control parameter which is set to 16 obtained from preliminary tests.

Stage 6. Segregated Signal Reconstruction

To segregate a target speech signal, the segregation filter $G(\lambda,k)$ is applied to both the left and right observed signals as follows

$$S^{R}_{target}(\lambda,k)=G(\lambda,k)\cdot X_R(\lambda,k);\ S^{L}_{target}(\lambda,k)=G(\lambda,k)\cdot X_L(\lambda,k) \tag{12}$$

The target speech signals are then obtained by overlapp-adding the windowed outputs

$$s^{R}_{target}(i)=\sum_{\lambda} IFFT\left\{S^{R}_{target}(\lambda,k)\right\}_{trunc}\cdot w(i-\lambda\cdot\frac{N}{2}) \tag{13}$$

$$s^{L}_{target}(i)=\sum_{\lambda} IFFT\left\{S^{L}_{target}(\lambda,k)\right\}_{trunc}\cdot w(i-\lambda\cdot\frac{N}{2}) \tag{14}$$

where $IFFT\{\cdot\}_{trunc}$ means that the output signal of IFFT must be truncated (due to the zero-padding) to restore the original size N of the signal.

3. FDBM Implementation on SBCs

For synchronization of audio and visual information, a latency between input and output signal must be negligibly short in order to reduce the chance to be detected by a user. In many real-time audio applications, low latency is very important. Especially in the case of hearing assistance system, latency may cause an echo in the perceived sound when users is listening to their own voice [6]. Thus, reducing latency is one of major issue on FDBM implementation.

Zero padding of FFT is used on SBC to implement FDBM to maintain a high-frequency resolution using a short frame length. As well known that, in a real-time application, a frame length refers to a buffer size where if it is too short we can suffer

crackles and audio dropouts. Thus, performing real-time processing on recent devices, such as ARM-based single board computer (SBC), needs to be discussed.

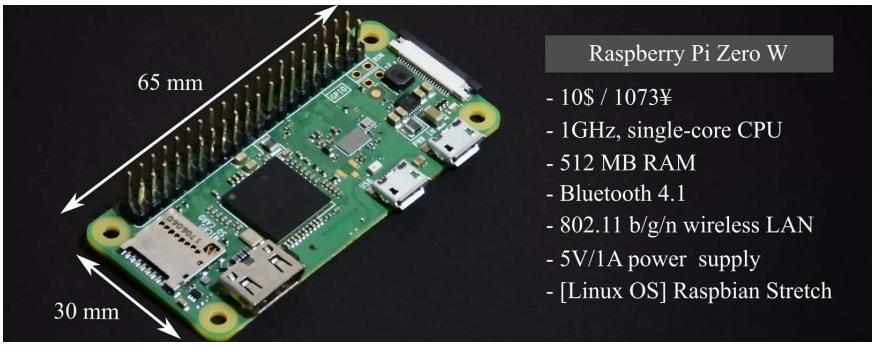

Figure 2. ARM-based single board computer (SBC)

Figure 3(a) shows the simulation condition where the target speaker is located at 0, and the interferer at -45°. The sampling frequency is 16 kHz, and SNR is set to 0 dB. Fifty male and fifty female speech signals were used as target speaker and interferer, respectively. The aim of this simulation is to investigate whether or not using zero-padding allows the use of a shorter buffer size (< 32 ms) and can still maintain the quality of segregated sounds, evaluated by Perceptual Evaluation of Speech Quality (PESQ). Figure 3(b) shows the mean of PESQ scores as a function of frame length based on the simulation condition shown in Fig. 3(a). It can be seen that a shorter frame length could still be used where a similar PESQ score to the original FDBM was obtained when 16ms < N < 32ms.

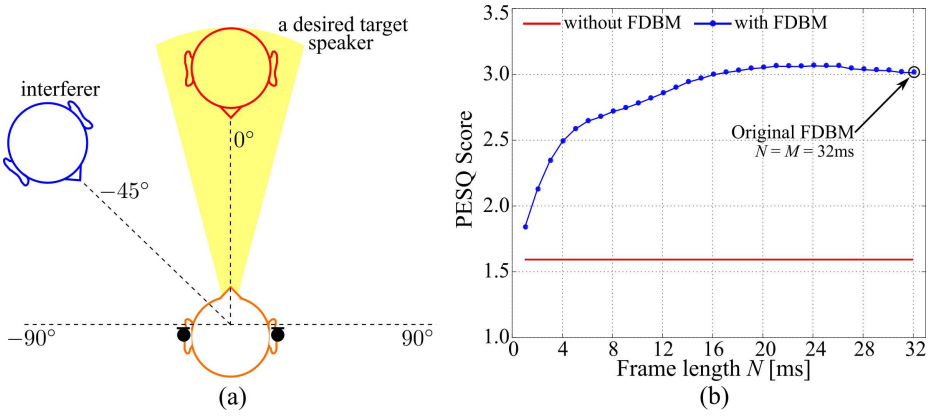

Figure 3. Simulation condition (SNR = 0 dB) and obtained PESQ against N

Hardware:

Here the hearing assistance device was assembled with available consumer hardware for about US$73 in total as shown in Fig. 4. The signal processing unit to which interfaces (i) an Andrea SB-205B – binaural microphone equipped earphones – is (ii) SBC: Raspberry Pi Zero W. This SBC runs the Raspbian, Debian-based OS, stored on (iii) a microSD card. It requires a power supply of 5V/1 A so that a lithium polymer (LiPo) battery: (iv) 3.7V 1300 mAH is equipped with (v) a step-up DC-DC converter. (vi) A little LiPoly charger is used to charge the battery via USB mini-B connector. (vii) A case from LEGO® was also made to cover all components.

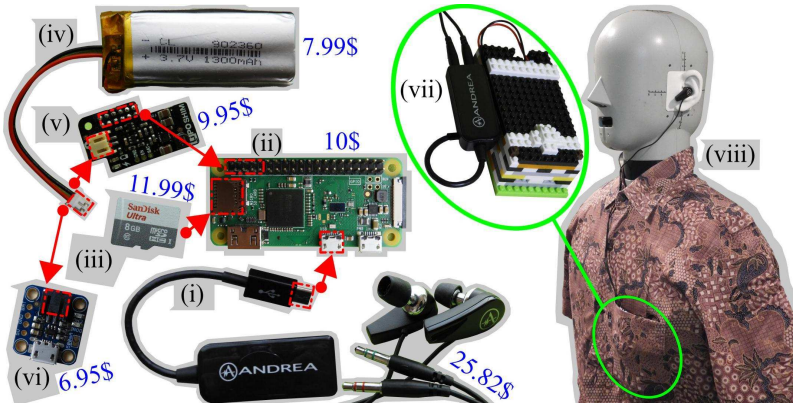

Figure 4. A hearing assistance system built with consumer hardware

Software:

The Python is the main language we used to run the system in real-time. Dummy head on right hand side of Fig. 4 shows an example of how to use the prototype on (viii) a dummy head. The list of hardware and software is available on our Github (https://github.com/shiinoandra/OpenFDBM).

4. Evaluation of FDBM on SBC

Directivity Pattern Measurement:

Figure 5 illustrates the procedure of the directivity pattern measurement as a function of direction. A dummy head wearing ear microphones/earphones was placed on the turntable and rotated every 5 degrees while five female and five male speech signals were produced by a BOSE 101VM loudspeaker located 1.4 m from the center of the dummy head. The prototype hearing device is running at sampling frequency of 16 kHz which we considered to be suitable for the Raspberry Pi Zero W to run the FDBM smoothly. When it was running, the computer used about 0.23A which can last over 3 hours when using a 1300-mAH battery. To control the device, an Android application so-called “OpenFDBM” has been developed using available open-source software.

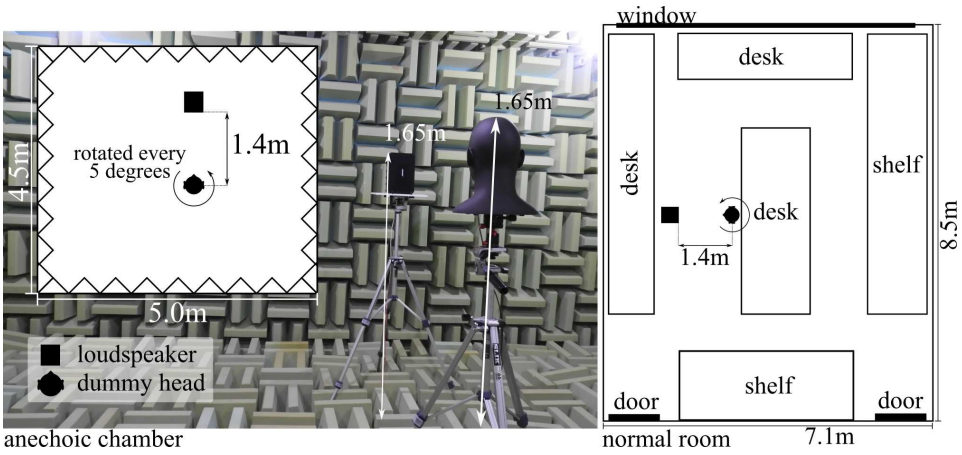

Figure 5. Schematic diagram to measure directivity pattern in an anechoic chamber and a normal laboratory room with a reverberation time (RT60) of 300 ms

Figure 6 shows the directional characteristics of the hearing device. It can be seen that the device running the FDBM enables a user to focus on a speaker in a certain direction, which is now only available for three directions: front, left and right sides (see a demo video https://youtu.be/FxcLsXWcGg0). In addition, "Front" mode does not attenuate the speech signals from the back (see Fig. 6). This phenomenon occurs because of a "front–back" confusion problem when using two microphones. A video example demonstrating how the "OpenFDBM" software is controlling the device in real-time is available at https://youtu.be/avdnHMR2AR4.

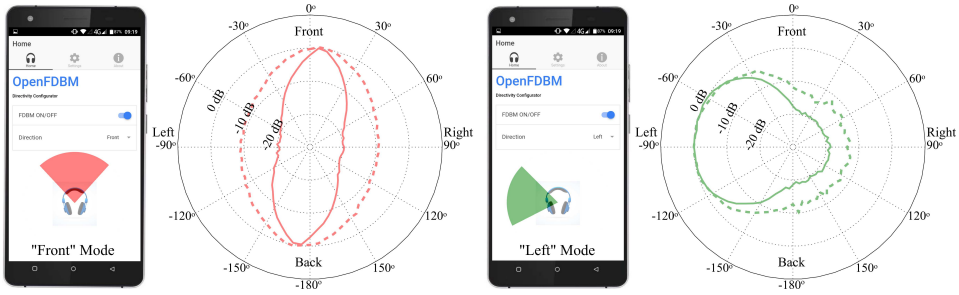

Figure 6. Directional characteristics of the Raspberry Pi hearing assistance device evaluated in an anechoic chamber (solid line) and a normal room (dashed line)

5. Conclusion

Real-time binaural speech segregation system using single board computer (SBC) has been successfully implemented. An "open-source" Raspberry Pi hearing assistance

device is assembled with available consumer hardware for about US$73 in total. Implemented FDBM allows to focus user's attention on a target speaker in a particular direction. An android application so-called "OpenFDBM" has also been developed to help users to choose the direction of a speech they want to focus on. Source codes including instructions to build the system are available on GitHub and demonstration video of this system is also available online on YouTube.

References

1. H. Nakashima, Y. Chisaki, Y. Usagawa, M. Ebata, *Frequency domain binaural model based on interaural phase and level differences*, Acoustical Science and Technology, **24** (2003) 172 – 174.
2. Y. Chisaki, K. Matsuo, K. Hagiwara, H. Nakashima, T. Usagawa, *Real-time processing using the frequency domain binaural model*, Applied Acoustics, **68** (2007) 923 – 938.
3. I. W. Selesnick, *Short-time fourier transform and its inverse*, 2009. (In: url http://eeweb.poly.edu/iselesni/EL713/STFT/stft_inverse.pdf visited on 06/24/2019).
4. N. Hiruma, R. Kouyama, H. Nakashima, Y. Fujisaka, *Low delay wind noise cancellation for binaural hearing aids*. INTER-NOISE, (2016) 4844 – 4854.
5. MIT-Media-Lab, *HRTF Measurements of a KEMAR dummy-head microphone*, 2019. (https://sound.media.mit.edu/resources/KEMAR.html visited on 06/24/2019).
6. J. Agnew, J. M. Thornton, *Just noticeable and objectionable group delays in digital hearing aids*, J. the American Academy of Audiology, **11** (2000) 330 – 336.