On Performance of Two-Sensor Sound Separation Methods Including Binaural Processors

Bagus Tris Atmaja[1,2], Yoshifumi Chisaki[1], Tsuyoshi Usagawa[1], Dhany Arifianto[2] *
[1]Kumamoto University, Japan
[2]Institut Teknologi Sepuluh Nopember, Indonesia

Abstract

Human beings have binaural inputs to separate and localize sound sources. Those two functions of binaural hearing can not be easily transformed to the computational methods. In this paper, three conventional methods to separate target signal from interfering noise are compared. Those methods include a binaural model, an independent component analysis (ICA) and a time-frequency masking applied to ICA. Performances were compared by means of spectrograms as well as coherence.

1 Introduction

As the cocktail party problem defines, human being has ability to focus on target sound than others. The left and right ears work simultaneously to localize and segregate the mixed sounds with the neuro-processing mechanism. How the human auditory processing solved this problem is not yet cleared until new. However, many researchers have proposed various methods to approach the cocktail party phenomena. The approaches can be built through modeling of human auditory system, statistical-mathematical approach and psychological-computational modeling. This paper compares three methods in sound separation task.

2 Methods

Following three methods which can be used for signal enhancement task using binaural inputs are examined in this paper.

2.1 Binarual Model

Most models of human binaural hearing are derived from binaural cues i.e. ITD (interaural time difference) and ILD (interaural level difference). The binaural model examined here is derived from phase difference in frequency domain to estimate the ITD as described in [1]. The binaural model is referred to Phase Difference Channel Weighting (PDCW) and it is described as follows. At first, binaural signals are observed by two microphones are transformed into time-frequency domain by means of STFT. Then ITD is estimated through comparison of binaural signals at each frequency. The time-frequency mask is identified in time-frequency domain at which ITDs are closed to the ones corresponding to the target source. After the gammatone channel weighting is applied, the resynthesis process is performed by means of inverse STFT and overlap-add method. Although the details explanation of PDCW algorithm can be found in [1]. Key of this method is how to identify the specific time-frequency bin which is dominated by target source. PDCW makes the binary decision whether the time-frequency bin belongs to target source or not based on the ITD for

*2-39-1 Kurokami, Kumamoto 860-8555
(bagus@hicc, chisaki@cs, tuie@cs).hicc.cs.kumamoto-u.ac.jp, dhany@ep.its.ac.id

each time-frequency bins.

2.2 ICA

Let $\mathbf{s(n)}$ be sampled signal of sound signal, n denotes the discrete time index. In convolutive mixture problem, let N be statistically mutually independent sources $\boldsymbol{s(n)} = [s_1(n), \ldots, s_N(n)]^T$ and M mixture observations $\boldsymbol{x(n)} = [x_1(n), \ldots, x_M(n)]^T$ are given by

$$\mathbf{x}(n) = \sum_{k=0}^{K} \mathbf{A}(k)\mathbf{s}(n-k), \qquad (1)$$

where $\{\mathbf{A}(k)\}$ is a sequence of $M \times N$ matrices. Sound separation is a problem to estimate the sound signal from its mixture observations without prior information of the mixing process. In causal finite impulse response (FIR) filter, separation process can be casted into,

$$\mathbf{y}(n) = \sum_{l=0}^{L} \mathbf{W}(l)\mathbf{x}(n-l) \qquad (2)$$

where $\mathbf{y}(n) = [y_n(n), \ldots, y_m(n)]^T$ are the independent estimate of each source $\mathbf{s}(n)$. $\mathbf{W}$ is $N \times M$ separation matrix, in which the quality of separation process depends on this variable. In this paper, FastICA algorithm introduced by Aapo Hyarinen [2] is used. FastICA algorithm uses non-gaussianity measure based on negentropy. This algorithm is formulated by fixed-point iteration, and has the same formulation derived from Newton's method . Rule of weighting factor $\mathbf{W}$ in this algorithm given by,

$$w^{+} = E\left\{xg\left(w^T x\right)\right\} - E\left\{g^{'}\left(w^T x\right)\right\} w \qquad (3)$$

$$w = \frac{w^{+}}{\|w^{+}\|} \qquad (4)$$

Where g is derivative of contrast function to approach non-gaussianity and norm w is used to check if the new w is convergence, if not, the algorithm will go back to calculate w^{+}.

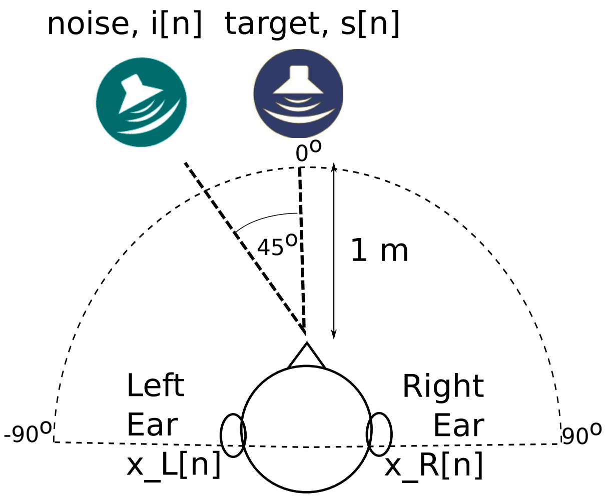


Fig 1: Location of target and interfering noise

2.3 ICA with Time-Frequency Masking

Like the binaural model described in section 2.1, the ICA with time-frequency masking is proposed by Wang et al. who were motivated by human auditory phenomenon in which a sound is rendered by louder sound within critical band[3]. The mask $M(n,k)$ in time-frequency domain is expressed as

$$m\left(n,k\right) = \begin{cases} 1 & if\ S_1\left(n,k\right) - S_2\left(n,k\right) > \theta \\ 0 & otherwise \end{cases} \qquad (5)$$

where n and k stand for indexes of time and frequency, and $S_1(n,k)$ and $S_2(n,k)$ stand for spectral components for the target and interference signals. Because $M(n,k)$ has binary weights, this method can be called as ICA with binary masking. The threshold θ is set to 0 corresponding 0 dB.

3 Experiment Design

Experiment is performed under the setup shown in Fig. 2. Two loudspeakers are located in the direction of 0^o and -45^o of a head and torso simulator (HATS) as shown in Fig. 2. The target signal is female speech where white noise is used for interference. The level of interference is set to the same

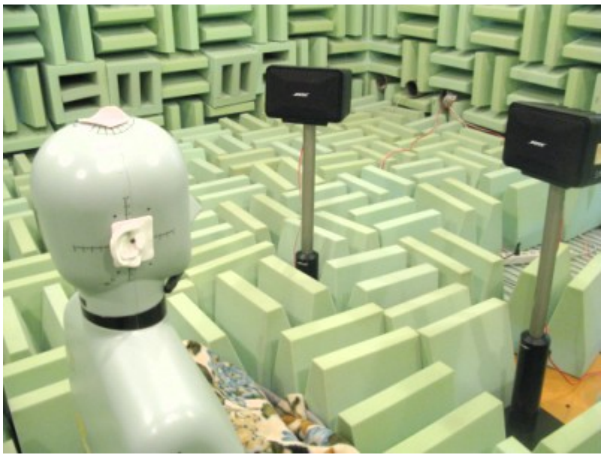


Fig 2: Experiment in Anechoic Room

level of target signal at the loudspeakers' positions. Although observed signals by microphones of HATS are recorded in 44.1 kHz sampling with 16 bit resolution of PCM, signals are down-sampled at 16 kHz because a major interest as the target is speech.

4 Results of Experiment

Figure 3 shows the observed signals as waveforms and the spectrograms of those signals. The order of plots in those figures is, from top, target signal, observed signal at left ear position, one at right, enhanced signal by the binaural model (PDCW) mentioned in 2.1, one by Fast ICA in 2.2, and one by ICA with binary masking (ICABM) in 2.3. Note that ICABM method utilizes the binary mask adopted from [4]. Also Table 1 shows the averaged coherences between the target signal and each of three enhanced signals. In this table, results of simulation are also provided for the comparison. According to preliminary subjective evaluation, noise in mid frequency range is reduced by PDCW and the sound quality of PDCW seems to be among FastICA and ICABM. FastICA provide the best according to the coherence criterion. ICABM provides fair performance according to the spectrogram and coherence criterion. Although the waveform of ICABM seems to be similar to target signal, it has low coherence. ICABM minimizes interfering noise and remains target signal, but the sound quality is degraded as can be seen by its spectrogram as in Fig. 3.

Table 1: Objective evaluation in coherence

Method	Simulation	Experiment
$PDCW$	0.542	0.283
$FastICA$	0.669	0.351
$ICA + BM$	0.539	0.277

5 Conclusion

This paper presents a comparison among three methods on sound separation algorithms based on binaural inputs. The result of performance is shown as averaged coherence for simulation results as well as experimental results. Based on coherence measurement, the FastICA algorithm work better than PDCW and ICABM. As the future works, the various conditions of signals not only arrangements as well as types of signals need to be examined in order to make clear the characteristics. Beside three methods examined in this paper, it is necessary to compare other methods such as [6] which is ready to be implemented as [7].

References

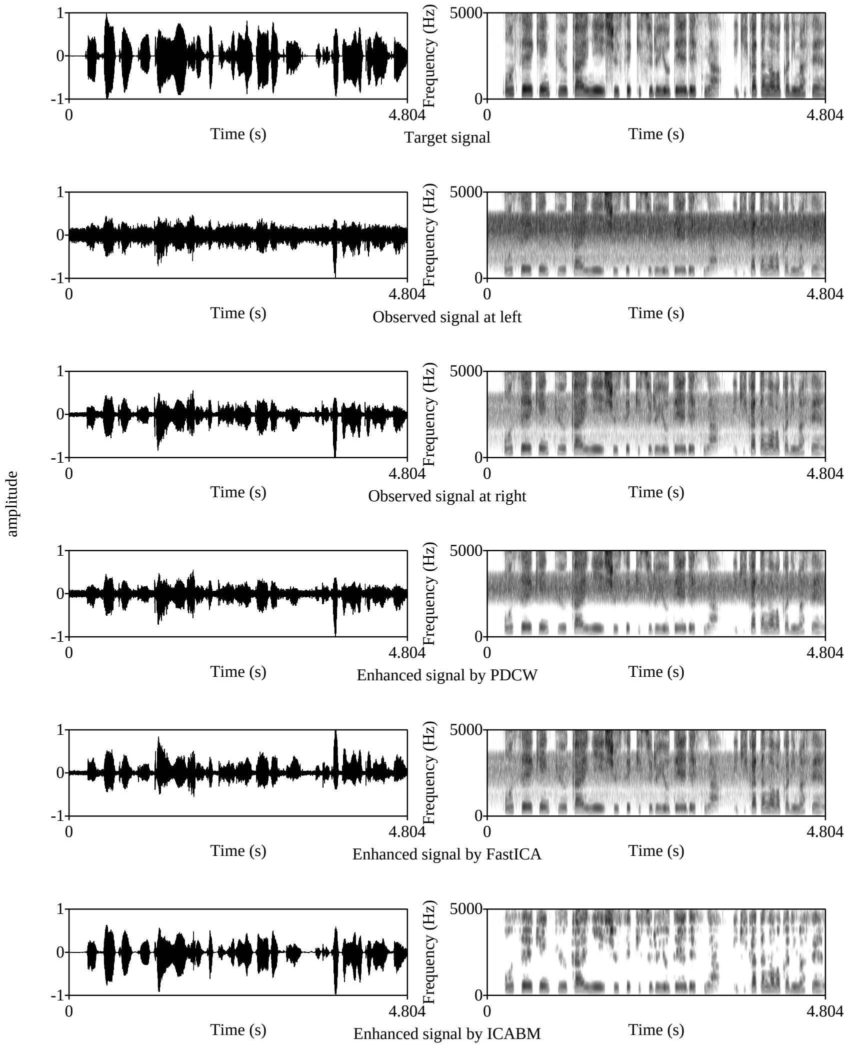
[1] Chanwoo Kim et al., INTERSPEECH 2009, pp. 2495-2498.

[2] Aapo Hyvarinen, IEEE Trans. on Neural Networks, 1999, Vol. 10(03), pp. 626-634.

[3] DeLiang Wang and Guy J. Brown, "Computatinal Auditory Scene Analysis: Principles, Algorithms and Applications," IEEE Press, 2006.

[4] M.S Pedersen et al., IEEE Trans. on Neural Networks, Vol. 10(20), pp. 1-18.

[5] http://sound.media.mit.edu/ica-bench/code/headmix.m



Target signal

Observed signal at left

Observed signal at right

Enhanced signal by PDCW

Enhanced signal by FastICA

Enhanced signal by ICABM

(a) Waveforms

(b) Spectrograms

Fig 3: Signals and its spectrogram in experiment

[6] Yoshifumi Chisaki et al. "Concurrent Speech Signal Separation Based On Frequency Domain Binaural Model," IWAENC 2003.

[7] Bagus Tris Atmaja and Dhany Arifianto, ICACSIS, 2009, pp. 259-263