A comparative study of sound sources separation by independent component analysis and binaural model

B T Atmaja, D Arifianto
Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
E-mail: {bagus,dhany}@ep.its.ac.id

Abstract. Humans' auditory system can separate mixed sounds based on their sources easily. However, mimicking this ability by computer algorithm is not an easy task. Some approaches have been developed, particularly based on the statistical approach and binaural modeling. From statistical methods, independent component analysis (ICA) grows fast to mimics sound separation and localization by human auditory processing. On the other side, mathematical modeling to model binaural hearing has been built block by block. This paper is a comparative study of both approaches, a statistical method represented by FastICA and binaural modeling represented by the frequency domain binaural model. The task is to mimic how to binaural processing works to separate sound sources. The result of the comparison was given by the perceptual evaluation of speech quality (PESQ) and Itakura-Saito (IS) distortion measurement. PESQ scores ICA method obtains better performance than the binaural model while, in contrast, IS scores the binaural model better than ICA.

1. Introduction
Binaural processing is the ability of humans auditory processing to separate and localize sound sources. This human intelligence solves the cocktail party problem to make selective listening, a capability to focus on one sound while listening to other sound sources. Across the multidisciplinary area, scientists and researchers developed some methods to mimics this human auditory processing. Two of the rigorous and actively developed methods were independent component analysis (ICA) and the binaural model (BM). ICA has been developed from a statistical approach, while the binaural model aimed to model human auditory processing block by block from input to output sounds.

While ICA exploits each sound source’s independence and utilizes it, the binaural model utilized the interaural phase difference (IPD) and interaural level difference (ILD) to get estimated sound sources. This paper compares those two methods side by side to study the advantages and disadvantages of both methods. Knowing each model’s advantages and disadvantages are useful to improve and redesign the system for a better sound separation quality. The quality or performance of sound source separation can be measured in several metrics, including PESQ (perceptual evaluation speech quality) [1] and spectral distortion measurement [2, 3].

Comparing several sound source separation algorithms has been conducted in many ways. In the previous work, we compared statistical approaches [4] and its a potential application for speech enhancement [5] while others proposed binaural model based on frequency domain...
(FDBM) to separate mixed sounds [6]. This paper compares the statistical approach with the binaural model to extend the previous evaluation.

The organization of this paper can be divided as follows. First, the cocktail party problem is introduced, and two approaches, i.e., ICA and FDBM are explained to solve the problem. How to obtain the data via an experiment is presented in section four, and how to measure the separation results is shown in section afterward. Finally, the results are discussed in the succeeding section. The summary of this research and a challenging opportunity for the next research is suggested in Conclusions.

2. Problem Statement
Source separation in the acoustics area came from the cocktail party problem, i.e., an event where multi speakers talk together. In this problem, human hearing can solve easily from who and where the sounds were emitted, even in a noisy environment like in a cocktail party. Both right and left ears work simultaneously to localize and segregate the mixed sounds with the neural processing mechanism. How the human auditory processing solved this problem is not clear until. Some models have been proposed, including auditory scene analysis (ASA), attention model, and binaural hearing.

![Figure 1. BSS Problem [7]](image)

Figure 1 shows an illustration of a simple cocktail party problem; two speakers talk simultaneously recorded by two microphones. Each microphone examines each speech signal at a different level and time; therefore, both two microphones received different unrelated signals. The sound sources (male and female) are independent speakers, resulting in independent signals on microphones. This unrelated information captured by two microphones leads to a strategy to separate individual signals from a mixture of voices. How to obtain estimated original signals and their locations are the goal of sound separation and localization problem. ICA and FDBM are two different approaches that are studied and compared in this paper.

The first method to solve the BSS problem presented in this paper is an independent component analysis, which separates a set of mixed signals into independent source signals. The second method is called the binaural model from the frequency domain, which exploits the differences of interaural time difference (ITD) and interaural level difference (ILD) received by two microphones. A brief explanation of these two methods is given below.

3. Independent Component Analysis (ICA)
ICA is one of solutions to solve the cocktail party problem by exploiting statistically independent of source signals. Let $s(n)$ be sampled signal of sound signal, $n$ denotes the discrete time index. In convolutive mixture problem, let $N$ be statistically mutually independent sources $s(n) = [s_1(n), \ldots, s_N(n)]^T$ and $M$ mixture observations $x(n) = [x_1(n), \ldots, x_M(n)]^T$ are given by

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

(1)
where \( \{ A(k) \} \) 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,

\[
y(n) = \sum_{l=0}^{L} W(l)x(n-l)
\]  

(2)

where \( y(n) = [y_n(n), \ldots, y_m(n)]^T \) are the independent estimate of each source \( s(n) \). \( W \) 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 [8] 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 Newtons method. Rule of weighting factor \( W \) in this algorithm given by,

\[
w^+ = E \left\{ xg \left( w^T x \right) \right\} - E \left\{ g' \left( w^T x \right) \right\} w
\]

(3)

\[
w = \frac{w^+}{\|w^+\|}
\]

(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 \( w^+ \).

Based on [9], FasICA with time-frequency masking is used in this paper. Binary mask using an ideal binary mask is motivated by the human auditory phenomenon in which a sound is rendered by louder sound within a critical band. The mask \( m(n,k) \) in the time-frequency domain is expressed as

\[
m(n,k) = \begin{cases} 
1 & \text{if } S_1(n,k) - S_2(n,k) > \theta \\
0 & \text{otherwise}
\end{cases}
\]

(5)

where \( n \) and \( k \) stand for indexes of time and frequency; \( S_1(n,k) \) 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 \( \theta \) is set to 0, corresponding to 0 dB.

4. Binaural Model

Most models of human binaural hearing are derived from binaural cues, i.e., ITD (inter-aural time difference) and ILD (inter-aural level difference). The binaural model examined here is derived from a phase difference in the frequency domain to estimate the ITD. The binaural model used in this paper is referred to as frequency domain binaural model (FDBM) developed by Usagawa et al. in the HICC Lab, Kumamoto University, Japan.

FDBM consist of some blocks as follow,

- FFT
- DoA estimation by IPD
- DoA estimation by ILD
- DoA estimation for frequency components
- DoA estimation for sound sources
- Segregation filter
- IFFT

The signals received on both canals, i.e., left and right, are in the time domain. For fast computation, this time-domain signals in discrete $x(n)$ are transformed to the frequency domain ($x(k)$) signal by using a fast Fourier transform method. As shown in figure 3, from FFT data, it can be obtained both ILD and IPD to estimate direction of arrival (DoA) of the frequency components and then estimate the direction of sound sources.

DoA estimation by IPD mainly from lower frequency bands. This estimation can be defined as follows,

$$C_{lr}(k) = L(xk)R(k)^*$$  \hspace{1cm} (6)

where * denotes a complex conjugate. The IPD is obtained by using cross spectrum.

The segregation filter, $m$, is obtained by using DoA information, azimuth and elevation. The segregated signal in the frequency domain then is inverse transformed back into the time domain by using inverse FFT (IFFT).

5. Method

Data presented in this paper was obtained from direct measurement in the anechoic chamber of HICC Lab, Kumamoto University. The first step to obtaining sound data reflected what we hear in the left, and right ears are obtained from the recording process by dummy head B&K (built-in microphones) connected to Roland R-44 recorder. The target speaker is female speech while interference is male speech; both are in the Japanese language. The experiment condition is shown in figure 4. The target speaker is located in $0^\circ$ of azimuth (in the front) while the interference speaker is located in $30^\circ$ (right side).
The next step is to input the data into ICA and FDBM methods for comparison and analysis.

6. Objective Measures
In this section, two objective measurements are presented to evaluate sound separation results from the data.

6.1. PESQ
Perceptual evaluation of speech quality (PESQ) is an objective evaluation for sound quality measurement proposed by ITU-T Recommendation P.862. PESQ takes into account human perceptual and psycho-acoustic models to generate results like the mean opinion score (MOS) derived from the human listener. It also used a cognitive model beside the perceptual model to measure the processed sound to the clean speech sound. The perceptual model process transforms the original and processed signals into the interaural representation based on
perceptual frequency (Bark) and loudness (Sone). The estimated subjective MOS is given by the cognitive model evaluating the difference between the original and processed signals. The PESQ score of the input signal is obtained from the original signal, while the output signal is the estimated signal. The range of the PESQ score is from 0.5 to 4.5, with 4.5 being the condition that the processed signal is exactly the same as the original signal.

6.2. Spectral Distortion
Spectral distortion shows how far the distortion of each frequency of the clean signal and after the separation process. The estimated signal might have attenuation or amplification from the original one. The measure of spectral distortion presented in this paper used the Itakura-Saito (IS) distance, which is shown in the equation below.

\[
D_{IS}(P(\omega), \hat{P}(\omega)) = \frac{1}{2\pi} \int \left[ \frac{P(\omega)}{\hat{P}(\omega)} \log \frac{P(\omega)}{\hat{P}(\omega)} - 1 \right] d\omega
\]  

(7)

The smaller value of IS distance shows the better quality of separation; the higher value represents the more distortion on each frequency after separation process compared to the original signal.

7. Results and Discussion
The separation results in waveform and spectrogram from both ICA and binaural model method are presented in figure 5 while the score of objective evaluation using PESQ and IS distance was shown in the table 1.

Table 1 shows the result of the separation of both ICA and FDBM methods from the same condition. The data obtained from 0 degree of the target signal (female speech) and interference (male speech) at 20 degree of azimuth. The signal to interference ratio was 25 dB SIR (signal to interference ratio). The PESQ score resulted from ICA using FastICA and a binary mask is 2.69, classified as a fair quality.

Table 1 shows that ICA performed a better performance in the PESQ score. PESQ is standard in speech quality measurement and is already used in industry, phone manufacture, network equipment, and telecom operators. It is recommended by International Telecommunication Union (ITU) for narrow-band telephone networks and speech codecs.

The result of the objective measure by IS distance show the opposite of PESQ for the first data. The higher quality of the estimated signal should have the lower IS distance score. However, the result shown in table 1 shows the higher PESQ, the lower the IS distance score. Compared to the spectrogram (figure 5), estimated signal from FDBM shows the more similar spectrum to original target signal. ICA’s estimated signal shows lost information in low frequency compared to the original spectrum of the target signal. By listening to the estimated sound directly, ICA’s estimated signal is perceptually better than the estimated signal by FDBM. Therefore, the objective measure by PESQ is more realistic and accepted as a standard quality measure in telecommunication.
While the first data (table 1 and figure 5 shows sound data from target speaker (female speech) and male interference in fixed SNR, Table 2 result of PESQ score of separation by ICA and FDBM in various dB SNR (white noise). In the first row, the target signal and noise are in the same power (0 dB SNR), and we increase the difference power between the target signal and noise from 0 to 25 dB by 5 dB steps. The more difference power between target and noise, the better the performance result shown by the PESQ score. Sound separation method by using ICA resulted in a higher PESQ score than FDBM, which is similar to the first data without noise variation.

Table 3 shows performance evaluation of separation result in IS distance score. The lowest score in IS distance is obtained by FDBM in which target signal and white noise has different power of 5 dB SNR. IS distance is a perceptual difference between original signal’s spectrum and the estimated signal’s spectrum. As previously mentioned, the score of IS distance is not as informative as PESQ score. The higher difference of power between target and noise should provide a better quality of separation (indicated by lower IS score), but it is not shown by IS distance. In table 3, the result is contradictive, the higher dB SNR, the higher IS distance score, which is come from the same data used to compute the PESQ score.

This paper shows two different methods to separate sound sources. ICA method using
Table 2. PESQ score in various SNR

| dB SNR | ICA | FDBM |
|--------|-----|------|
| 0 dB   | 1.10| 0.76 |
| 5 dB   | 1.67| 0.95 |
| 10 dB  | 2.08| 1.17 |
| 15 dB  | 2.17| 1.48 |
| 20 dB  | 2.35| 1.79 |
| 25 dB  | 2.60| 1.99 |

Table 3. IS distance score in various SNR

| dB SNR | ICA  | FDBM  |
|--------|------|-------|
| 0 dB   | 66.52| 9.63  |
| 5 dB   | 39.18| 9.5   |
| 10 dB  | 45.12| 10.50 |
| 15 dB  | 49.42| 13.82 |
| 20 dB  | 47.74| 19.86 |
| 25 dB  | 47.40| 25.08 |

FastICA and binary mask shows better performance by the PESQ score. FDBM, although it shows the lower PESQ score, it keeps a similar spectrum to the original target signal. The future improvisation on finding the better sound separation method should provide good quality sound both in perceptual and spectrum analysis with minimum information lost.

8. Conclusions

In this paper, we evaluated two different approaches to tackle the sound source separation problem by a computer algorithm. Two different metrics, PESQ and IS distance, shows a contradictive result. ICA shows a better result on PESQ score, while FDBM has a better performance on IS distance (lower distance). Using these two metrics, it is difficult to judge which one is better. A third evaluation has been performed using subjective listening and spectrogram plots evaluation. These subjective evaluations show ICA obtained perceptually better sound (confirming higher PESQ score) while FDBM obtained a more similar spectrogram to the original speech than (fast) ICA (confirming the IS distance measurements). Future research can be directed to obtain more consistent measures to judge the performance of sound separation algorithm, e.g., an algorithm that obtains good performance on PESQ, IS distance, and spectrogram plots.

References

[1] Rix A W, Beerends J G, Hollier M P and Hekstra A P 2001 Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) vol 2 (IEEE) pp 749–752
[2] Quackenbush S R, Barnwell T P and Clements M A 1988 Objective measures of speech quality (Prentice Hall)
[3] Juang B H 1984 AT&T Bell Laboratories technical journal 63 1477–1498
[4] Atmaja B T, Chisaki Y, Usagawa T and Arifianto D 2011
[5] Atmaja B T, Arifianto D, Chisaki Y and Usagawa T  Signal enhancement by using sound separation methods based on binaural inputs Basic Science Int. Conf pp 1–5

[6] Chisaki Y, Nakanishi T, Nakashima H and Usagawa T 2003 Concurrent speech signal separation based on frequency domain binaural model International Workshop on Acoustic Echo and Noise Control (IWAENC2003)

[7] 2007 Blind source separation via ica: Introduction and background OpenStax-CNX Web site, http://cnx.org/content/m15702/1.1/

[8] Hyvarinen A 1999 Fast ica for noisy data using gaussian moments 1999 IEEE International Symposium on Circuits and Systems (ISCAS) vol 5 (IEEE) pp 57–61

[9] Pedersen M S, Wang D, Larsen J and Kjems U 2005 Overcomplete blind source separation by combining ica and binary time-frequency masking 2005 IEEE Workshop on Machine Learning for Signal Processing (IEEE) pp 15–20