**THESIS TF2325**

# ON SOURCE SIGNAL SEGREGATION BASED ON BINAURAL INPUTS

Bagus Tris Atmaja
2410 201 006

Supervisor
Dr. Dhany Arifianto
Prof. Tsuyoshi Usagawa

Master Program
Department of Engineering Physics
Faculty of Industrial Technology
Institut Teknologi Sepuluh Nopember Surabaya
2012

This thesis was submitted to
Master of Engineering
at
Institut Teknologi Sepuluh Nopember

By :
Bagus Tris Atmaja
NRP. 2410 201 006

Thesis Defense : 6 February 2012
Period of Graduation : 10 March 2012

Approved By :

| | |
|---|---|
| 1. Supervisor | Dr. Dhany Arifianto |
| NIP : | 19731007 199802 1 001 |
| 2. Co-Supervisor | Prof. Tsuyoshi Usagawa |
| | Kumamoto University, Japan |
| 3. Examiner I | Dr. Bambang L. Widjiantoro |
| NIP : | 19690507 199512 1 001 |
| 4. Examiner II | Dr. Doty Dewi Risanti |
| NIP : | 19740903 199802 2 001 |
| 5.Examiner III | Dr. Gunawan Nugroho |
| NIP : | 19771127 200212 1 002 |

Director of Post Graduate Program

Prof. Dr. Ir. Adi Soeprijanto, MT
NIP. 19640405 199002 1 001

# Acknowledgments

On this page, I would like thank to many people helped me to this milestone.

- Alhamdulillahirobbil'alamin, all praise to Allah, Lord of this world, this nature and author of all whole things.
- Thank to my parent, my sisters, younger and older and all my family.
- Thank to Dr. Dhany Arifianto and Usagawa-Sensei as the supervisor of my thesis and my research. I feel like standing on the shoulder of two big giants to have supervisors like you.
- For the conducive condition in HICC Usagawa-Chisaki Lab., KU and Acoustic Lab, ITS, you make laboratory as it should be.
- To all my friend in post graduate Engineering Physics ITS, Kumamoto University and in over the world.
- To open source community, the developer of Ubuntu, Latex/Lyx, GNU Octave, Gedit, recordMydesktop, Gnome, Audacity, Inkscape, Libreoffice, moc player and others. You keep this world free of charge.
- To all who I cannot mention one by one, Thanks.

Kumamoto-shi, 25 January 2012
B.T. Atmaja

# ON SOURCE SIGNAL SEGREGATION BASED ON BINAURAL INPUTS

| | |
|---|---|
| By | : Bagus Tris Atmaja |
| Student Identify Number | : 2410 201 006 |
| Supervisor | : Dr. Dhany Arifianto |
| | Prof. Tsuyoshi Usagawa |

## ABSTRACT

Sound separation is interest problem in psychological and computational science. How the human auditory processing solved this problem is not yet cleared until now. This function of binaural hearing can not be easily transformed to the computational methods. Independent Component Analysis (ICA) is one of methods to solve blind source separation (BSS) and the fast algorithm is implemented by FastICA. Binary time-frequency masking is another approach using filtering methods to obtain basis signals. The combination of ICA and FastICA with binary mask can estimates sources signals from binaural inputs. This research evaluate systematically various methods: ICA, ICA with binary mask, binaural model with PDCW, FastICA and FastICA with binary mask. Performances were compared by means of coherence measurement and PESQ. The results show that proposed method has the best performance on male speech interference and white noise interference at -20 dB and -10 dB of SIR.

Key words: ICA, FastICA, Binaural Model, Binary Mask, Binaural Inputs, Coherence, PESQ

# Pemisahan Sumber-Sumber Suara Tercampur Dari Binaural Inputs

| | |
|---|---|
| Nama | : Bagus Tris Atmaja |
| NRP | : 2410 201 006 |
| Pembimbing | : Dr. Dhany Arifianto<br>Prof. Tsuyoshi Usagawa |

## ABSTRAK

Pemisahan sumber (*source separation*) merupakan teknik pemisahan beberapa sumber sinyal (suara) tercampur menjadi komponen-komponen (sumber-sumber) penyusun campuran tersebut. Pemisahan dengan banyak sensor, seperti *large-scale microphone array* umumnya menghasilkan kualitas pemisahan sinyal yang lebih bagus untuk memisahkan sinyal-sinyal suara tercampur daripada menggunakan mikrofon dengan jumlah lebih sedikit. Hal ini dikarenakan secara matematik akan lebih mudah untuk merubah data ukuran besar menjadi lebih kecil daripada sebaliknya. Sedangkan manusia hanya memiliki dua sensor pendengaran saja (*binaural sensor*s), telinga, namun mampu memisahkan sinyal suara dari berbagai macam sumber bunyi yang berbeda. Pada penelitian ini, dilakukan pemisahan sumber-sumber bunyi tercampur hanya dari dua mikrofon saja sebagai sensor. (BSS) adalah metode pemisahan sumber berdasarkan sinyal campuran terukur sensor saja, dimana sensor yang digunakan biasanya lebih dari satu (array sensor). Proses pemisahan tersebut dilakukan dengan memanfaatkan kebebasan antar sumber, dimana sumber satu bebas secara statistik terhadap sumber lainnya (*Independent Component Analysis*, ICA) dan komputasi secara cepat direalisasikan dengan algoritma FastICA. Metode ICA dan FastICA dievaluasi pada penelitian ini serta dikombinasikan juga dengan binay mask dan metode binaural model dengan teknik *phase difference channel weighting* (PDCW). Beberapa kondisi yang dievaluasi adalah variasi jumlah sumber bunyi, variasi perbedaan SNR dan variasi frekuensi pencuplikan. Evaluasi performansi secara obyektif diberikan dengan kriteria koherensi dan PESQ. Hasil penelitian ini menunjukkan metode FastICA dengan binary mask yang diusulkan memperoleh hasil terbaik pada beda SIR -20 dB dan -10 dB serta pada interference *male speech*.

Kata kunci: ICA, Binary Mask, Binaural Model, Pemisahan sumber bunyi, Koherensi, PESQ

# Contents

Validation Sheet **iii**

Acknowledgments **v**

Abstract **vii**

Abstract in Indonesian **ix**

Table of Contents **xi**

List of Figures **xiii**

List of Tables **xv**

I INTRODUCTION **1**

I.1 Background 1

I.2 Problem Statement 2

I.3 Research Purposes 2

I.4 Research Benefits 2

I.5 Thesis Outline 2

II SOUND SOURCES SEGREGATION **5**

II.1 Independent Component Analysis (ICA) 5

II.2 ICA with Binary Mask 7

II.3 *Binaural Model* 8

II.4 FastICA 9

II.5 FastICA with Binary Mask 11

II.6 Objective Evaluation 11

II.6.1 Coherence Criterion 11

II.6.2 PESQ Score 12

III SIMULATION AND EXPERIMENT **15**

III.1 Simulation 15

III.2 Experiment 16

IV RESULTS AND DISCUSSION **19**

IV.1 Results of Simulation as Opposed to Experiment Data 19

IV.2 Types of Interferences 22

IV.3 Effects of Various SIR 24

IV.4 Effects of Various SNR 25

IV.5 Effects of Various Sampling Frequency 26

V CONCLUSION **29**

Bibliography **31**

List of Publications **35**

# List of Figures

II.1 Block diagram of ICA with binary mask (ICABM)[1]: showing the process of ICABM to obtain basis signals from the output of ICA 8

II.2 Block diagram PDCW 9

II.3 Block diagram of FastICA with binary mask (FastICABM) adapted from [2] 12

II.4 The overview of PESQ [3] 13

III.1 Block diagram of simulation and separation system 16

III.2 Location of target signal and interference 17

III.3 Screenshot of experiment 17

III.4 Diagram of multi-channel recording in anechoic room. Remarks: (1) Bř uel & Kjær Head and Torso Simulator, Type 4128, (2) BNC connector (microphones to amplifier), (3) Roland R-44 Recorder (wiring using connector TRS), (4) Speaker Bose Type 101VM (Sound sources: female speech, male speech and white noise), (5) Amplifier Bose Type 1706II, (6) PC (Macbook MB403JA). 18

IV.1 Waveform and spectrogram from simulation data. The top one is original signal, the following two signals are observed signals and the last three signals are enhanced signal from some methods. 20

IV.2 Waveform and spectrogram from experiment data. From the top are target signal, observed signal at left ear position, one at right, enhanced signal by the PDCW, FastICA and ICABM. 21

IV.3 Comparison of coherence criterion for various dB SNR (target vs additive noise) 26

IV.4 Comparison of PESQ score for various dB SNR (target vs additive noise) 26

IV.5 Comparison chart of coherence criterion in various sampling frequencies 27

# List of Tables

IV.1 Comparison of averaged coherence score from simulation and experiment data 22

IV.2 Separation task of target signal and white noise interference (SIR= 20 dB, SNR= 10 dB) 22

IV.3 Separation task of target signal and male interference (SIR= 20 dB and 4 dB, SNR= 10 dB) 23

IV.4 Separation task of target signal with male and white noise interference (SIR= 4 dB and 20 dB, SNR= 10 dB) 23

IV.5 Comparison of coherence criterion from separation task in various SIR 24

IV.6 Comparison of PESQ score from separation task in various SIR 25

# CHAPTER I

# INTRODUCTION

## I.1 Background

Signal enhancement is an important thing in which implemented speech recognition, hearing aids and telecommunication. In speech recognition, it will be difficult to recognize sound with mixed noise. It applies also in hearing aids and telecommunication. People with hearing aids require better sound quality to easily understand communication to other person. In telecommunication, the signal enhancement using source separation is needed to separate sound of speaker from other noises.

The enhancement task can be done through separating the target sound and the interference sounds. This sound separation method is an interest problem in psychological and computational science motivated by human auditory processing. The way of human audiotory processing remanis unknown. The function of binaural hearing can not be easily transformed to the computational methods. Independent Component Analysis (ICA) is one of existing methods to solve blind source separation (BSS) problem, This method separates mixed sources without prior knowledge (blind) of source and mixing process. It exploits the statistically independent nature of sources. FastICA implemented fixed point algorithm to obtain faster convergence speed for ICA [4]. Another way to obtain basis signal from mixed signal is by using binary time-frequency masks. This thesis evaluate some methods in source separation from the point of view from computational science, modeling and mathematic-statistics. There are five methods to be evaluated in this thesis : ICA, ICABM, PDCW, FastICA and FastICABM. Goal of this research is to obtain enhanced signal target compared mixed signal from the left and right ears. Therefore, enhanced signals were measured by means of coherence criterion and PESQ score to obtain performance comparison.

## I.2 Problem Statement

There are two problem statement addressed in this research :

1. Comparing some methods including proposed method on source separation problem for signal enhancement task based on binaural inputs.
2. Measuring the objective evaluation by means of coherence and PESQ.

## I.3 Research Purposes

The purposes of this research are to solve the problem statement in previous section. First, some methods on two-sensor source separation are evaluated including ICA, ICABM, PDCW, FastICA and FastICABM. Then, enhanced signal was evaluated by means of coherence criterion and PESQ score. This research want to know which method has the fair quality and close to human source separation ability. However, it will be difficult because there are two different objective evaluation and which one is the best depend on its application.

## I.4 Research Benefits

The results of this thesis should be beneficial for source separation, signal enhancement, sound engineering or in general signal processing. For a special purpose, this research is can be useful for investigating of machinery fault diagnostic tools in Acoustic Laboratory, ITS, in which one of the problem is to reduce the number of sensors to separate machine sounds. Since this research uses almost all of speech data instead of machine sound data, for machine sound data further evaluation is needed.

## I.5 Thesis Outline

This thesis consist of five chapters and summarized as follows. The first chapter (introduction) describes briefly reason why for conducting this research. It includes thesis background, problem statements, purposes, benefits and thesis outline. Chapter two of this thesis explained the theories used in this research. They are five methods used in this research: ICA, ICABM, PDCW, FastICA and ICABM. The theory of objective evaluation which consist of coherence criterion and PESQ calculation is also outlined. The third chapter explain simulation and experiment methodology which describes step-by-step way to conduct this research and also the equipment used along with their set up. Chapter four present result and discussion. Separation result from different tasks are shown. The last chapter is conclusion which states important findings of this thesis based on the conducted research.

# CHAPTER II

# SOUND SOURCES SEGREGATION

This chapter explains the theory and concept used in research as well as the method to measure the objective evaluation. Some methods used in this research are ICA, ICA with binary mask, binaural model using phase difference channel weighting (PDCW), FastICA and FastICA with binary mask. The objective evaluation is given by means of coherence criterion and perceptual evaluation of speech quality (PESQ) score.

## II.1 Independent Component Analysis (ICA)

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 $\boldsymbol{s(n)} = [s_1(n),\ldots,s_N(n)]^T$ and $M$ mixture observations $\boldsymbol{x(n)} = [x_1(n),\ldots,x_M(n)]^T$ are given by

$$x_i(n) = a_{i1}s_1(t) + a_{i2}s_2(t) + .... + a_{in}s_N(n) \tag{II.1}$$

each of $x_i = x_i(n)$ consist of samples in number of $N_s = f_sT$. $T$ is in second or millisecond and $f_s$ is sampling frequency. Eq. II.1 can be written in matrices-vector form as follows,

$$x(n) = As(n) + v(n) \tag{II.2}$$

or,

$$x_i(n) = \sum_{j=1}^{N} a_{ij}s_j(n) \tag{II.3}$$

Eq. II.3 is named as model of independent component analysis (ICA) in general form. The independent component ($s$) is latent variable, in which can not be observed directly. The mixing matrices, A, is also not known, only measured variable ($x$) is known. The component of $A$ and $s$ can be estimated from measured signal $x$. Then, to estimate the observed signal $s$, one can use new variable namely $y$ and formulated as follows,

$$y(n) = W(x(n) - v(n)) \tag{II.4}$$

where $W = A^{-1}$

The main problem in ICA is to find $W$ filter where $W$ is inverse of mixing matrix $A$. The more proper choice of $W$ the better quality of separation results. As explained in [2], the use of ICA methods is to separate mixing sound of industrial machinery. The paper results reported of ICA implementation in machine sound separation although the performance was reported not good enough [5]. However, it proved that ICA method may work for machine separation problem.

Maximum likelihood concept can be used to determine separation matrix $W = (w_1, w_2, ..., w_N)^T$ for ICA method [6]. Likelihood from a model is probability function of set of data which is used as model parameter. Model of mixed signal, $x$ was characterized by mixing matrix $A$, and density of $p$ source can be defined as the following

$$L(W) = \log \prod_{n=1}^{N} p_x(X(n)) \tag{II.5}$$

where $p_x(X)$ is pdf (probability density function) of mixed signal and it is correlated with the pdf or source signal with the following formulation

$$p_x(X) = |J|\, p_s(S) \tag{II.6}$$

where $|J|$is absolute value of Jacobian matrix from equation II.4. Jacobian matrix is determinant of partial derivative,

$$J = det \begin{pmatrix} \frac{\partial s_1}{\partial x_1} & \cdots & \frac{\partial s_1}{\partial s_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_M}{\partial x_1} & \cdots & \frac{\partial s_N}{\partial x_M} \end{pmatrix} \tag{II.7}$$

Hence,

$$J = det\, \mathbf{W} \tag{II.8}$$

Equation II.4 and II.8 can be included in equation II.5 so that the following formula can be obtained,

$$L(W) = \sum_{n=1}^{N} \log p_s(\mathbf{Wx}(n)) + T \log |\det \mathbf{W}| \tag{II.9}$$

If the sources are statistically independent then that equation become,

$$L(W) = \sum_{n=1}^{N} \sum_{i=1}^{M} \log p_i(\mathbf{W}_i \mathbf{x}(n)) + T \log |\det \mathbf{W}| \tag{II.10}$$

where $p_i(s_i)$ is density of $i-th$ sources. That equation can be solved using natural gradient or Newton and similar iteration.

## II.2 ICA with Binary Mask

The use of binary time-frequency mask in this research was motivated from the phenomena in human auditory system in which a sound is rendered by louder sound within critical band [7]. The term of “masking” here means weighting (filtering) the mixture, which is different from the same term used in psycho-acoustics where it means blocking the target sound by using acoustic interference[8]. The mask $M(n,k)$ in time-frequency domain is expressed as

$$m(n,k) = \begin{cases} 1 & if\, s_1(n,k) - s_2(n,k) > \theta \\ 0 & otherwise \end{cases} \tag{II.11}$$

where $n$ and $k$ are index of time and frequency. $s_1(n,k)$ and $s_2(n,k)$ define spectral components of the interference and target signal. Due to $m(n,k)$ has binary linear weighting function hence it is called as binary mask. Threshold $\theta$ is chosen to be 0 referred to 0 dB as stated in reference [7].

Combination of ICA and binary mask can be obtained by filtering the output of ICA with binary mask [1]. This method estimates basis signals from mixed signal in iterative way from two input channels. The number of outputs can be different depend on input signal and algorithm when working while the output of ICA or FastICA is the same as the number of inputs. Figure II.1 shows the diagram block of ICA with binary mask. It can be shown from that ICABM works in iterative way (indicate by loop arrows) to estimate sources signal and it also eventually saved the estimated signal as stereo wav sound.

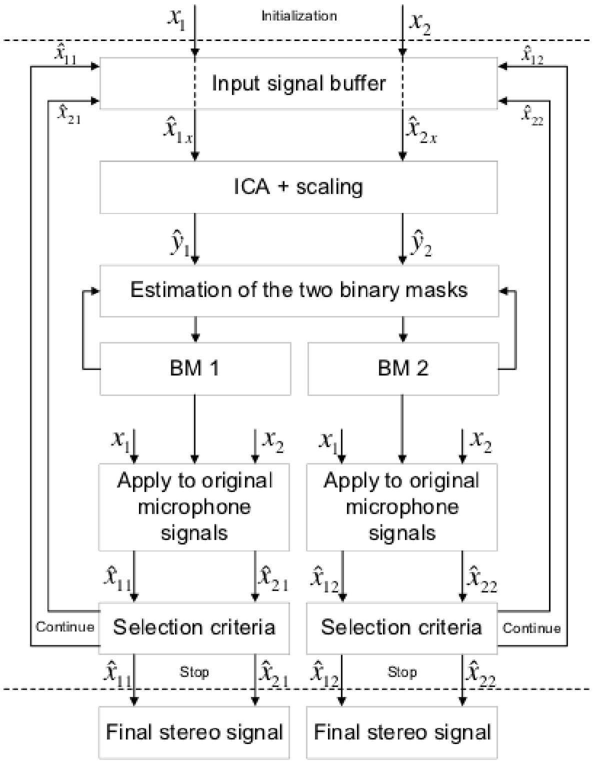


Figure II.1: Block diagram of ICA with binary mask (ICABM)[1]: showing the process of ICABM to obtain basis signals from the output of ICA

## II.3 *Binaural Model*

Binaural model is common approach to model human auditory system and generally it is derived from the concept of ILD (*interaural level difference*) and ITD *(interaural time difference*) between right ear and left ear. By exploiting ITD and ILD in different ways, some binaural model can work to separate and localize sound sources.

The binaural model examined here is derived from phase difference in frequency domain to estimate the ITD as described in [9]. 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, they are transformed into time-frequency domain by means of short time Fourier transform (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 (STIFT) and overlap-add method (OLA). Details explanation of

PDCW algorithm can be found in [9]. 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 each time-frequency bins. Figure II.2 shows the way PDCW work to separate sound sources and obtain enhanced target signal.

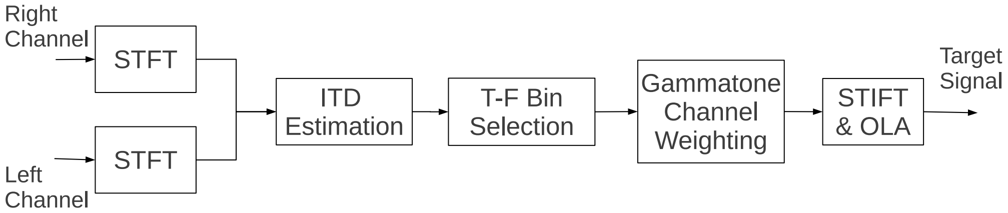


Figure II.2: Block diagram PDCW

Basic formulation of input signals in binaural model using PDCW method can be shown in the following equation. $x_L$is mixed signals entering left channel and $x_R$ is mixed signal at the right channel.

$$x_L[n] = \sum_{l=0}^{L} x_l[n], x_R[n] = \sum_{l=0}^{L} x_l[n - \delta(l)] \tag{II.12}$$

## II.4 FastICA

Fixed-point theorem stated that function $F$ minimum has one fixed value (at $x$ where $F(x) = x$), in some same condition, $F$ can be stated generally [6]. Fixed-point point theorem can be expanded to fixed-point iteration. Fixed point iteration the is formulated as follows,

$$x_{n+1} = g(x) \tag{II.13}$$

where $g$ is function obtained from the transformation of $x = g(x)$. The FastICA is based on a fixed-point iteration scheme for finding a maximum of the nongaussianity of $w^T x$ [6]. The FastICA learning rule is designed to find a direction, i.e. a unit vector $\mathbf{w}$ such that the projection $\mathbf{w}^T\mathbf{x}$ maximizes nongaussianity. Nongaussianity here is measured by the approximation of negentropy $J(\mathbf{w}^T\mathbf{x})$,

$$J(y) \approx \frac{1}{12} E\left\{y^3\right\}^2 + \frac{1}{48} kurt(y)^2 \tag{II.14}$$

The random variable $y$ is assumed to be zero mean and unit variance[6]. However, the validity of such approximations may be rather limited. In particular, these approximations suffer from the nonrobustness encountered with kurtosis [6].

This problem can be tackled by using new approximation based on the maximum-entropy principle:

$$J(y) \approx \sum_{i=l}^{p} k_i [E\{G_i(y)\} - E\{G_i(v)\}]^2 \tag{II.15}$$

where $k_i$ are some positive constants, and $\nu$ is a Gaussian variable of zero mean and unit variance (i.e., standardized). The variable y is assumed to be zero mean and unit variance, and the functions $G_i$ are some nonquadratic functions. It is noted that even in cases where this approximation is not very accurate, can be used to construct a measure of nongaussianity that is consistent in the sense that it is always non-negative, and equal to zero if y has a Gaussian distribution.

In the case where we use only one nonquadratic function G, the approximation becomes

$$J \propto [\{\{G(y)\} - E\{G(v)\}\}] \tag{II.16}$$

for practically any non-quadratic function G. This is clearly a generalization of the moment-based approximation in, if $y$ is symmetric taking $G(y) = y^4$, one then obtains exactly, i.e. a kurtosis-based approximation [6].

It is important that G that not grow too fast, one obtains more robust estimators. The following choices of G have proved very useful[6],

$$G_1(u) = \frac{1}{a_1} \log\cosh a_1 u \tag{II.17}$$

$$G_2 = -\exp\left(-\frac{u^2}{2}\right) \tag{II.18}$$

where $1 < a_1 < 2$ is constant variable. Equation II.17-18 can be simplified:

$$g_1(u) = \tanh(a_1 u) \tag{II.19}$$

$$g_2(u) = u \exp\left(-\frac{u^2}{2}\right) \tag{II.20}$$

For $a_1 = 1$, then FastICA algorithm for one unit can be started using random separating filter of $w$ according to the size of input matrix. Then, the next $w$

based on [4] can be obtained by using the following formulation,

$$w^{+} = E\left\{xg\left(w^{T}x\right)\right\} - E\left\{g^{'}\left(w^{T}x\right)\right\}w \tag{II.21}$$

$$w = \frac{w^{+}}{\|w^{+}\|} \tag{II.22}$$

where $g$ is contrast function from non-gaussianity approach and norm $w$ is to check whether the new $w$ is convergent or not. If it is not convergent, it will go back to calculate $w^{+}$. Convergent means the new $w$ and the old one at the same direction.

## II.5 FastICA with Binary Mask

FastICA with binary mask used in this research is similar to ICA with binary mask from [1] and as explained in section II.4. The reported ICA as shown in Fig. II.1 is changed to FastICA. Based on previous research [10], it can be concluded that output of FastICA result remains noise although it has the highest objective evaluation score and also close to the original signals. In the other side, the sound quality of enhanced signal by ICABM is fair enough but it is degraded and has low objective score measured by means of coherence criterion [10]. The combination of FastICA with binary mask may accommodate the advantages and disadvantages of those two methods.

The diagram block of FastICABM can be show in Fig. II.3. The figure shows the work of FastICABM in which it exploits FastICA output after scaling process and performing binary mask filtering in iteratively to obtain estimated signals. These signals are then as stereo signals.

## II.6 Objective Evaluation

### II.6.1 Coherence Criterion

Coherence is the measure how well a signal correlated to other signal. Coherence measurement is usually used to measure the quality of processed signal compared to original clean signal. Coherence can be formulated as follows,

$$C_{x,y}(k) = \left|\frac{Wxy(k)}{W_{xx}(k)W_{yy}(k)}\right| \tag{II.23}$$

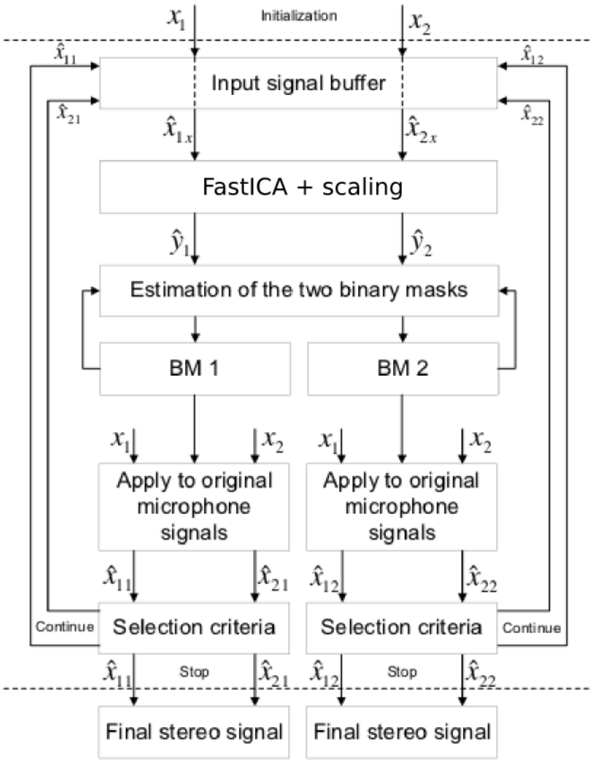


Figure II.3: Block diagram of FastICA with binary mask (FastICABM) adapted from [2]

where $W_{xy}(k)$ is cross spectrum from $x(n)$ and $y(n)$, $W_{xx}(k)$ is power spectrum from $x(n)$ and $W_{yy}(k)$ is power spectrum from $y(k)$.

The magnitude of a coherence is a function of frequency. The coherence value can be scored by its averaged and it is varied from $0-1$, where 0 indicates no correlation between two signals in frequency domain and 1 refer to processed signal being exactly same as original clean signal. The averaged coherence value can be obtained from the following formulation.

$$AveragedCoherence = \sqrt{\frac{C_{xy}C'_{xy}}{N_C}} \tag{II.24}$$

where $C'_{xy}$ is transpose of $C_{xy}$ and $N_c$ is length of vector $C_{xy}$.

### II.6.2 PESQ Score

Perceptual evaluation of speech quality (PESQ) is objective evaluation for sound quality measurement proposed by ITU-T Recommendation P.862 [3]. PESQ takes into account human perceptual and psycho-acoustic models to generate results like the mean opinion score (MOS) derived from human listener. It also used cognitive model beside 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. In the simulation, the reference signal, i.e. the original signal as shown in Fig. II.4, is always the target signal. The PESQ score of input signal is obtained when the processed signal shown in Fig. II.4 is $y$ or estimated signal. The range of the PESQ score is from $0.5 \sim 4.5$ with 4.5 being the condition that the processed signal is exactly the same as the original signal.

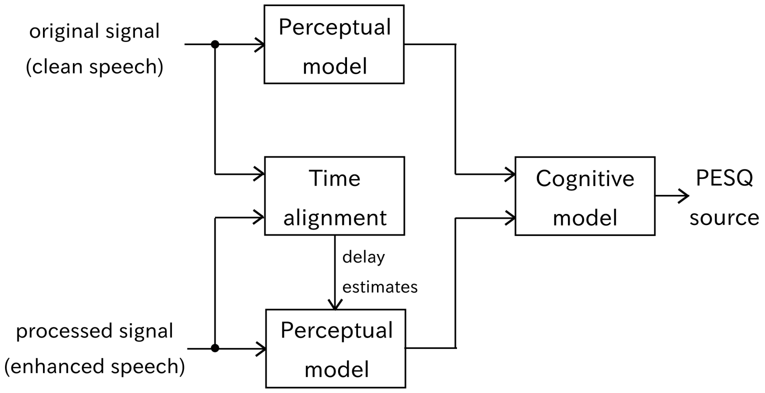


Figure II.4: The overview of PESQ [3]

# CHAPTER III

# SIMULATION AND EXPERIMENT

Most of data used in this research are from simulation. The experiment is performed to analyze the difference between them. Sound separation methods described in previous chapter are used to process data obtained from the following methods.

## III.1 Simulation

Simulation to obtain mixed sound data can be done by convolving source signal with head related transfer function. The observed signals from simulation can be defined as follow,

$$x_L(n) = \sum_i h_l(\theta_i) * s_i + n_l(n) = l_0(n) + l_1(n) + l_2(n) + \ldots + n_l(n), \qquad \text{(III.1)}$$

and

$$x_R(n) = \sum_i h_r(\theta_i) * s_i + n_r(n) = r_0(n) + r_1(n) + r_2(n) + \ldots + n_r(n), \qquad \text{(III.2)}$$

where $x_L$ and $x_R$ are observed signal at left and right, $h_l(\theta_i)$ and $h_r(\theta_i)$ are HRTFs from the direction $\theta_i$ and $s_i$ is sound from the $i-th$ direction. $l_0(n)$ and $r_0(n)$ are target signal observed at left and right, $l_1(n)$ and $r_1(n)$ are interference signals. The goal of separation system is to obtain better target signal symbolized with $y(n)$. The * symbol represent convolution operation. $n_l(n)$ and $n_r(n)$ represent additive noise at left and right. The noise was assumed from the hardware (wire, PC, recorder and microphones). Figure 3.1 shows the diagram block of simulation in computational method.

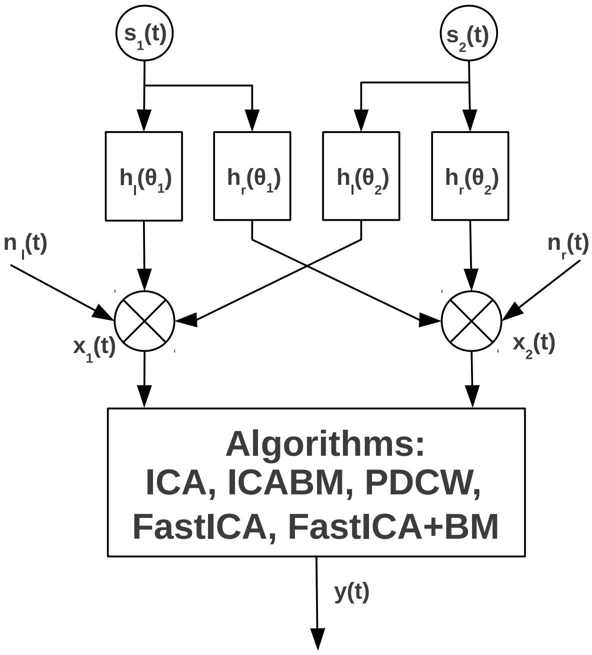


Figure III.1: Block diagram of simulation and separation system

## III.2 Experiment

For the experiment, sound sources were located in anechoic room at Usagawa-Chisaki Laboratory where recording process was undertaken to obtain sound data. The location and position of sound sources i.e target signal and interference for example can seen in figure III.2, which also represents for simulation data. In that figure, target signal was located at elevation 0° and azimuth $0°(0°, 0°)$ while interference noise was located at elevation 0° and azimuth $-45°$ $(0°, -45°)$. Screenshot of experiment in anechoic room can be seen in Figure 3.3. Figure III.3 show the realization in experiment set up using dummy head or head and torso simulator (HATS) Bṙuel & Kjær type 4128 with loud speaker as the sound sources. In the dummy head are exist small microphone located in ear of dummy head. The recording process was performed by contacting wire from microphone to amplifier then to recorder. For this purpose, Roland R-44 recorder was employed. Figure III.4 shows the block diagram of recording process.

The experiment used Japanese female speech as target signal as well as for simulation data. The interference signals were white noise and Japanese male speech. The sentence spelled by Japanese female target is /o-n-se-i-ke-n-kyu-ka-i-shu-s-se-ki-su-ru-yo-te-i-de-su-ga-i-ki-ka-ta-ga-wa-ka-ri-ma-se-n/ which means “I will attend speech seminar but I don’t know how to go there”. Sampling frequency used when record sound is 44100 Hz. While in computational process, 16000 Hz

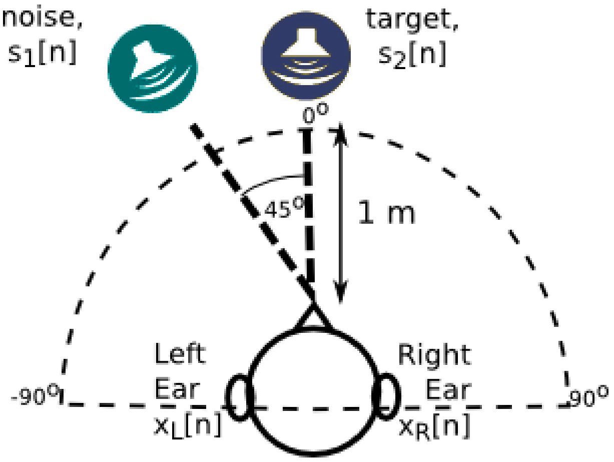


Figure III.2: Location of target signal and interference

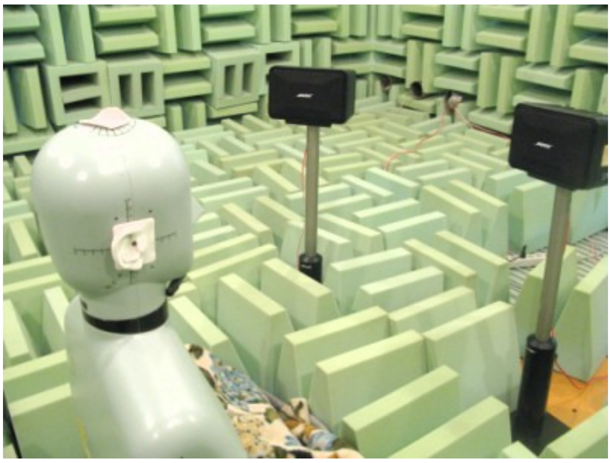


Figure III.3: Screenshot of experiment

was used because the main focus is human speech. Others sampling frequencies i.e 48000 Hz, 22050 Hz and 8000 Hz were also used to evaluate the impact of different sampling frequency in separation result.

As the result of recording process, sound data were obtained in form of two .wav (16 bits PCM) files for each condition which representing sound listened at left and right ears. Those two wav files will be used to extract target signal from interference noise and it is expected that target signal is free from interference sound after separation process. To accommodate other noises which can be avoided, two independent additive noises were added to simulation as shown in equation III.1 and III.2 and also in diagram block in Figure III.1. The SIR (Signal to Interference Ratio) for target and white noise is 20 dB while female target and male interference has 4 dB of SIR. The additive noise was also varied to evaluate the separation result as the function of additive noise. Block diagram of recording process in anechoic room, Usagawa-Chisaki laboratory can be seen in figure III.4.

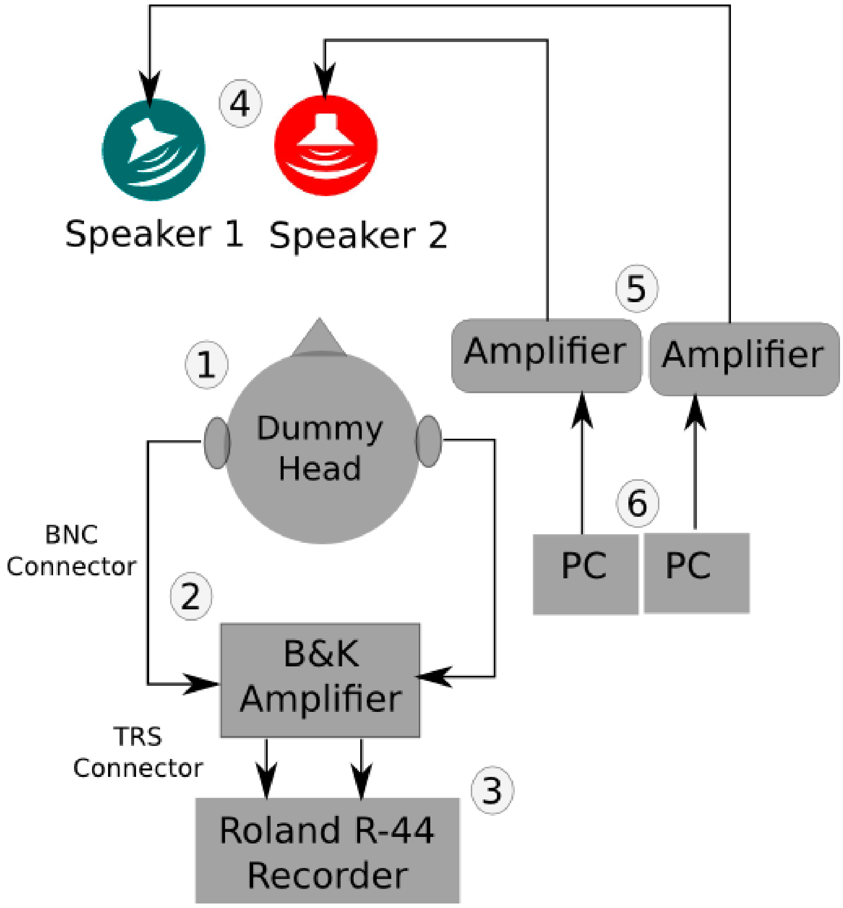


Figure III.4: Diagram of multi-channel recording in anechoic room. Remarks: (1) Bṙuel & Kjær Head and Torso Simulator, Type 4128, (2) BNC connector (microphones to amplifier), (3) Roland R-44 Recorder (wiring using connector TRS), (4) Speaker Bose Type 101VM (Sound sources: female speech, male speech and white noise), (5) Amplifier Bose Type 1706II, (6) PC (Macbook MB403JA).

# CHAPTER IV

# RESULTS AND DISCUSSION

This chapter presents the result of sound source segregation using some methods explained in Chapter II and methodology as explained in Chapter III. Some tasks and conditions are evaluated as well as objective evaluation by means of coherence criterion and PESQ score.

## IV.1 Results of Simulation as Opposed to Experiment Data

Figure 4.1 shows waveform and spectrogram of source segregation task from simulation data (section III.1) and Figure 4.2 shows waveform and spectrogram from experiment data (section III.2). Location and condition for these result were shown in FigureIII.2 where target signal is located in $(0°,0°)$ and interference signal in $(0°,-45°)$.

Data from simulation are combination of the following variation of variables.

- Sound sources: Japanese female speech (target signal), Japanese male speech, white noise (interference signals)
- Elevation : -10, 0, 10
- Azimuth : ±90, ±75, ±60, ±45, ±30, ±15, 0
- SIR (Target vs interference) = 20 dB, 10 dB, 0 dB, -10 dB, -20 dB
- SNR (Target vs additive noise) = 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB
- HRTF = MIT, Nagoya University
- Sampling Frequency = 48000, 44100, 22050, 16000, 8000 Hz

Not all data were examined in this research because the complexity and space. As example, the previous data as shown in Figure IV.1 is the data where target signal located in elevation of 0° and azimuth 0° while interference sound using white noise was located at elevation 0° and azimuth $-45°$. For the next, sound data located in position of elevation and azimuth was symbolized by $(elevation, azimuth)$.

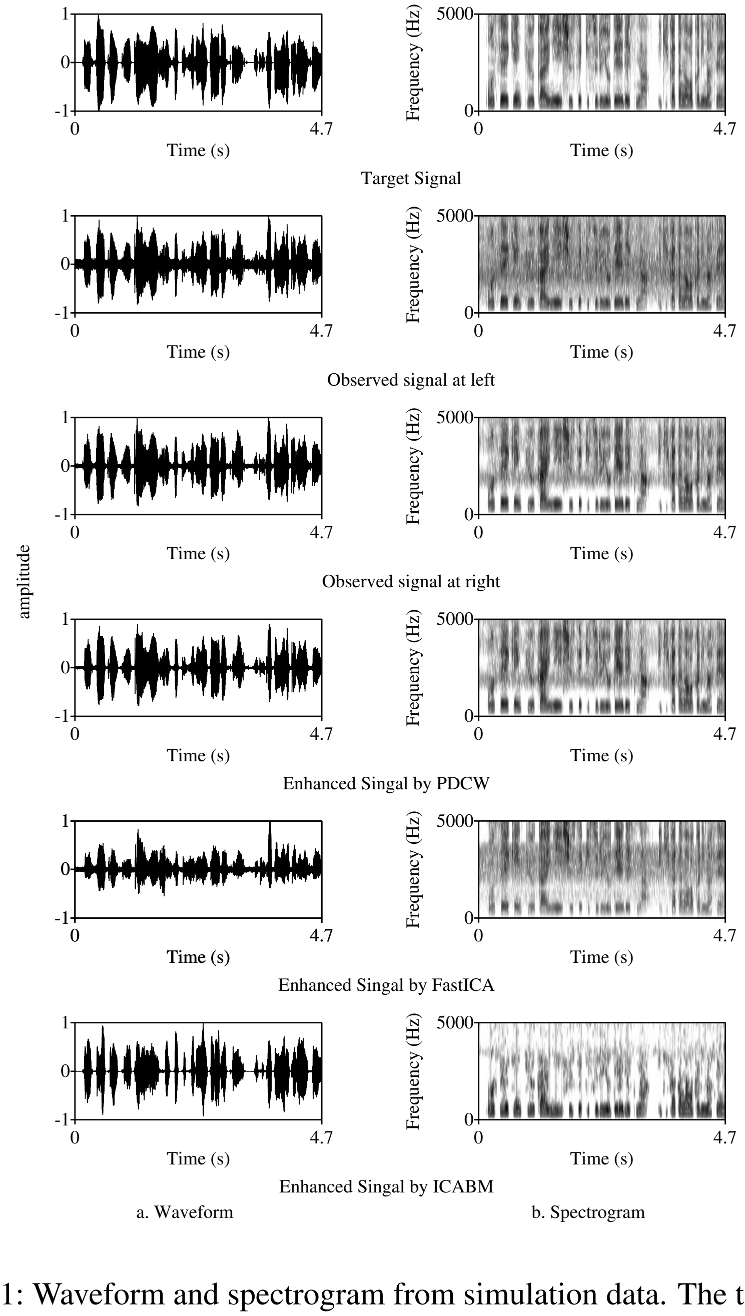
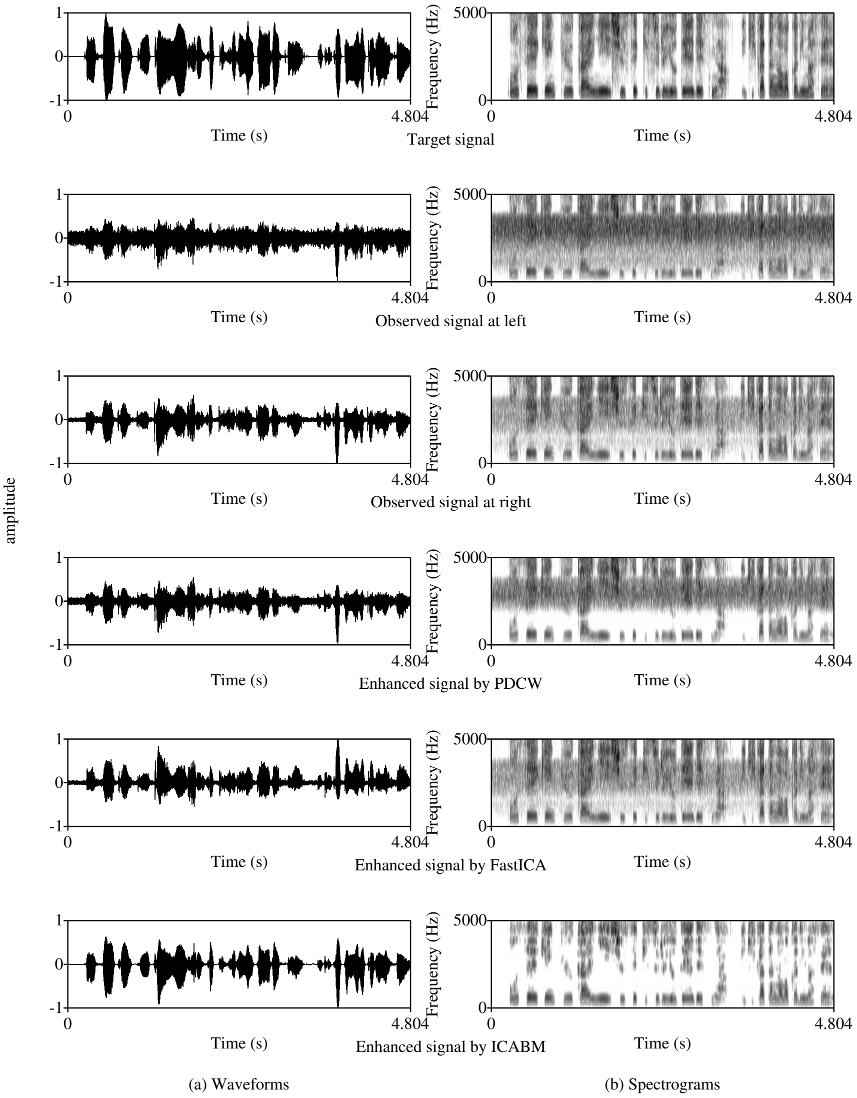


Figure IV.1: Waveform and spectrogram from simulation data. The top one is original signal, the following two signals are observed signals and the last three signals are enhanced signal from some methods.

Figure IV.2 illustrates waveform and spectrogram analysis of experiment data. It is seen that result of waveform from ICABM is close to original target signal as well as by listening enhanced sound. From spectrogram analysis, the FastICA is more similar to the original and it also supported by averaged coherence score in which enhanced signal from FastICA has the highest score among other methods. In other side, enhanced signal by ICABM has low coherence although it has similar sound and waveform. The respective experiment data were performed to separate target signal from interference noise. The results of separation task from experiment data was given in Figure 4.2.



(a) Waveforms

(b) Spectrograms

Figure IV.2: Waveform and spectrogram from experiment data. From the top are target signal, observed signal at left ear position, one at right, enhanced signal by the PDCW, FastICA and ICABM.

Table IV.1: Comparison of averaged coherence score from simulation and experiment data

| Methods | Simulation Data | Experiment Data |
|---|---|---|
| PDCW | 0.542 | 0.283 |
| FastICA | 0.669 | 0.351 |
| ICABM | 0.539 | 0.277 |

Table IV.1 shows the averaged coherence among the target signal and each of three enhanced signals. In this table, results of simulation were 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 provides the best separation result 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 degraded target signal. Its spectrogram indicates that sound quality is degraded as can be seen in Figure IV.2.

## IV.2 Types of Interferences

The first task was performed with two different types of interference. This can be casted into three conditions of separations based on types on interferences; separation of target signal from white noise interference, separation of target signals from male speech interference and separation of target signal from male speech and with noise interferences. The results of those conditions were compiled in Table IV.2, IV.3 and IV.4, respectively.

Table IV.2: Separation task of target signal and white noise interference (SIR= 20 dB, SNR= 10 dB)

| Methods | Coherence | PESQ |
|---|---|---|
| ICA | 0.72 | 1.94 |
| ICABM | 0.68 | 1.95 |
| PDCW | 0.58 | 1.91 |
| FastICA | 0.72 | 1.94 |
| FastICABM | 0.79 | 1.90 |

Result of separation task from where target signal was corrupted with white noise interference was shown in IV.2. The SIR (target vs interference) was set to 20 dB while SNR (target vs additive noise) was 10 dB. Noise was included to simulation to close the real environment. From this condition, the highest coherence score was obtained by FastICABM and the PESQ scores are almost similar of the five methods does not differ significantly.

Table IV.3: Separation task of target signal and male interference (SIR= 20 dB and 4 dB, SNR= 10 dB)

| Methods | Coherence | PESQ |
|---|---|---|
| ICA | 0.74 | 2.08 |
| ICABM | 0.72 | 2.5 |
| PDCW | 0.55 | 1.56 |
| FastICA | 0.73 | 2.08 |
| FastICABM | 0.72 | 2.46 |

The second task in analyzing effect of different interference was done by using male speech interference. Although the SIR between target signal and interference noise is only 4 dB, the result was fair enough by means of PESQ score (Table IV.3). PDCW has 1.562 of PESQ score while others exhibit PESQ score above 2.0. PDCW method has the lowest score in coherence criterion and PESQ score. The method previously used 4 cm of microphone distance [9], but in this research, simulation of binaural recording was performed using HRTF in which the distance of left and right ear is about 19 cm. Spatial aliasing might occurs in PDCW method and it reduced the separation quality of PDCW method [10].

Table IV.4: Separation task of target signal with male and white noise interference (SIR= 4 dB and 20 dB, SNR= 10 dB)

| Methods | Coherence | PESQ |
|---|---|---|
| ICA | 0.72 | 1.75 |
| ICABM | 0.68 | 2.02 |
| PDCW | 0.58 | 1.33 |
| FastICA | 0.72 | 1.75 |
| FastICABM | 0.72 | 2.01 |

The last condition of evaluating different types of interference was performed by corrupting female speech target with male and white noise simultaneously. The male speech was located in $(0°, 30°)$ and white noise interference was in $(0°, -30°)$. The results from that condition is listed in Table IV.4. The values of SIR and SNR were set to be same as the previous values i.e 20 dB and and 4 dB of SIR, and 10 dB of SNR. In this case, ICABM has the highest score of PESQ while the highest score in coherence was obtained by ICA and FastICA.

From this evaluation of different types of interference, it can be shown that signal enhancement by source separation method is better performed with male speech interference. To decide which method is the best, there are two choices of criterion, coherence and PESQ score. The coherence criterion is good for further application like speech recognition and the similar tasks, while PESQ is perceptually motivated from human hearing system, so it is suitable for perceptual application such as hearing aids and telecommunication.

## IV.3 Effects of Various SIR

The third task of separation problem presented in this paper is to evaluate the effect of various signal to interference ratio (SIR) on separation result by means of coherence criterion and PESQ score. White noise was chosen as interference signal located in $(0°, 30°)$ while target signal remains in $(0°, 0°)$. The separation result by means of coherence criterion are presented in Table IV.5 which shows all result from five methods with various SIR. However, when SIR between target signal and interference signal was changed to 0 dB, 10 dB and 20 dB, conventional ICA method exhibits the highest coherence score. Table IV.5 also indicates that ICA, PDCW, and FastICA methods has no significant change particularly at -20 dB, -10 dB 10 dB and 20 dB of SIR. It might be concluded that the algorithm cannot differentiate SIR from -20 dB and -10 dB, also 10 dB and 20 dB. In hearing science, the binary mask will work like in two tone suppression on certain threshold of dB SIR. In this research, the threshold can be predicted between -10 dB to 0 dB and 0 to 10 dB. It is also can be concluded that the result of computational method is similar and obey the result from psychological approach. However, further investigation are needed to check the results for human auditory behavior and experiment data.

Table IV.5: Comparison of coherence criterion from separation task in various SIR

| Methods | SIR | | | | |
|---|---|---|---|---|---|
| | -20 dB | -10 dB | 0 dB | 10 dB | 20 dB |
| ICA | 0.60 | 0.60 | 0.59 | 0.63 | 0.63 |
| ICABM | 0.60 | 0.60 | 0.39 | 0.33 | 0.33 |
| PDCW | 0.51 | 0.50 | 0.40 | 0.21 | 0.2 |
| FastICA | 0.60 | 0.59 | 0.59 | 0.63 | 0.63 |
| FastICABM | 0.63 | 0.61 | 0.32 | 0.42 | 0.47 |

Table IV.6 shows results of separation task in PESQ score for various SIR. Again, at -20 dB and -10 dB of SIR, FastICABM reveals the highest score, while at 0 dB highest PESQ score was obtained by ICABM and at 10 dB and 20 dB, conventional ICA has the highest value among others. The result of ICABM close to perceptual evaluation by listening the enhanced sound. In this case, PESQ score is more suitable when sound separation is designed for perceptual application such as in hearing aids and telecommunication.

Table IV.6: Comparison of PESQ score from separation task in various SIR

| Methods | SIR | | | | |
|---|---|---|---|---|---|
| | -20 dB | -10 dB | 0 dB | 10 dB | 20 dB |
| ICA | 1.18 | 1.18 | 1.18 | 1.38 | 1.38 |
| ICABM | 1.19 | 2.08 | 1.55 | 0.69 | 0.70 |
| PDCW | 1.17 | 1.17 | 1.19 | 0.99 | 1.00 |
| FastICA | 1.18 | 1.18 | 1.18 | 1.38 | 1.38 |
| FastICABM | 1.27 | 2.11 | 1.28 | 0.94 | 1.27 |

## IV.4 Effects of Various SNR

In the previous task of separation problem, condition of sound sources segregation were varied using various value of SIR while the value of SNR was fixed at 20 dB. This research takes into account the additive noises assumed from background noises or other noises from hardware which usually being excluded in other research[6, 2, 1].

Six different SNR values were evaluated from 0 dB to 25 dB with interval of 5 dB. The results show increasing objective evaluation score for both coherence and PESQ scores. As shown in Figure IV.3, at 0 dB of SNR, the averaged coherence criterion is 0.45 while the PESQ score at that condition is 1.04. The highest value was achieved at the highest dB SNR i.e 25 dB with averaged coherence value of 0.72 and PESQ score of 2.69. As opposed to of previous task in section IV.3, this task was performed with fixed value of signal to interference ratio (SIR) i.e 4 dB between female speech and male speech interference. The results show hat the additive noises affect the separation results indicating that the additive noises strongly affect the separation results of separation task. FastICABM algorithm was used to evaluate separation result of different SNR and the obtained PESQ score was the highest among all data and tasks.

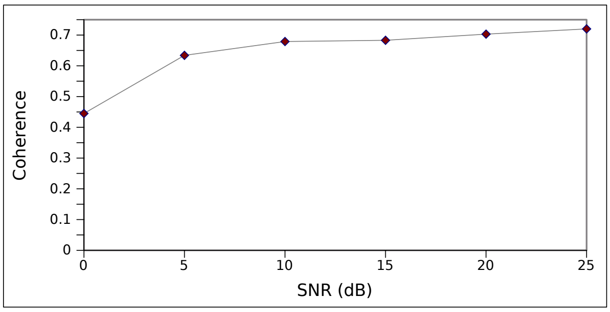


Figure IV.3: Comparison of coherence criterion for various dB SNR (target vs additive noise)

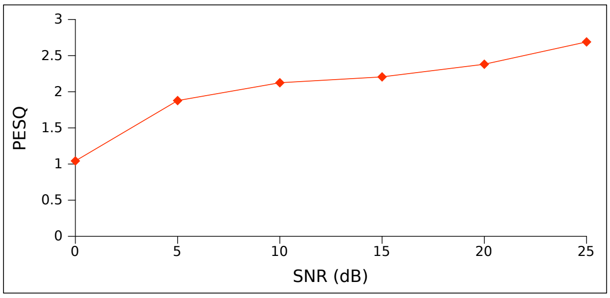


Figure IV.4: Comparison of PESQ score for various dB SNR (target vs additive noise)

## IV.5 Effects of Various Sampling Frequency

The last task of separation problem was performed by evaluating different sampling frequency to the sound sources and HRTF. The standard sound files from simulation was recorded in 16000 Hz is down sampled and up sampled to be 8000 Hz, 22050 Hz, 44100 Hz and and 48000 Hz. The HRTF which has 44100 Hz of sampling frequency is also down sampled and up sampled to those kinds of sampling frequency. The different sampling frequency data then was used later as input of ICA algorithm. The objective evaluation for this task was only given by coherence criterion because PESQ was designed for 8000 Hz and 16000 Hz of sampling frequency only. Other algorithms like PDCW also cannot processed data above 16000 sampling frequency, therefore conventional ICA method is used.

The highest objective evaluation value by means of coherence criterion was obtained at frequency of 16000 Hz as shown in Figure IV.5. That figure also shown the use of higher sampling frequency does not gives significantly improvement for this research. However, using 16000 Hz of sampling frequency in real time processing such as in telecommunication needs more effort and my slowdown the separation process. The calculation of cost function should be considered for real application. At the 22050 Hz of sampling frequency, the highest coherence criterion was 0.84 which is the highest value among all data.

Further investigation of higher sampling frequency should be undertaken by implementing additional steps in computational program. Such as in music data, the higher frequency will give the better sound quality.

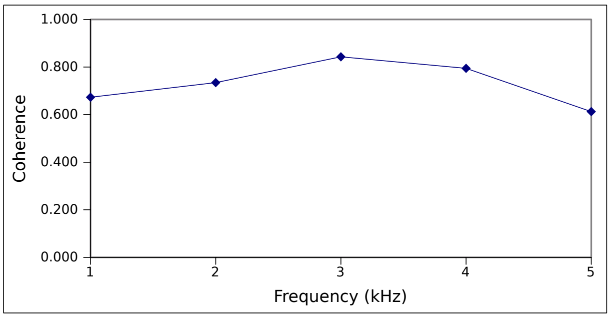


Figure IV.5: Comparison chart of coherence criterion in various sampling frequencies

# CHAPTER V

# CONCLUSION

This thesis evaluated some tasks and various conditions of source separation problems for signal enhancement. The first task evaluated the separation result from simulation data and experiment data showing that FastICA has the highest coherence criterion whereas ICABM gives fair sound quality although it is degraded. The second task was performed using simulation data to evaluate effect from kinds of interference signals. The different types of interference affected the separation result in which the higher objective evaluation was obtained by using male speech interference instead of white noise interference. Different SIR between target signal and interference signal also has impact on separation result. In -20 dB and -10 dB of SIR, FastICABM has the highest score of coherence and PESQ score. At 0 dB, PDCW obtain the better result of PESQ score and in 0 dB, 10 dB and 20 dB of SIR, conventional ICA method obtain the highest score of averaged coherence. The improvement of performance for both coherence and PESQ was obtained when increasing SNR to mixed sounds. The higher dB SNR between target signal and additive noise, the better separation result obtained. Noises actually cannot be avoided and it affected the separation result. Finally, the use of different sampling frequency gives the highest score of coherence criteria at 22050 Hz.

There are two objective evaluation to measure the performance of separation result besides the use of waveform and spectrogram analysis. To choose which objective is suitable for evaluation, either coherence criterion or PESQ score, is depend on the application. However, PESQ score is more suitable for perceptual application such as in hearing aids and telecommunication because it takes into psycho-acoustic model and cognitive model. For application such as speech recognition, speaker recognition, medical sound diagnostic and others related to this area, coherence measurement might be useful to measure the performance of separation results. Since for which, sound characteristic was extracted without perceptual needs.

For the future works, some methods used in this research need to be be examined for different conditions such as behavior of different sound sources elevation, location of sound sources in cone of confusion angle and other locations. Another issue is to reduce the gap between simulation and experiment result. By modeling sound separation in more detail and adapting human auditory processing mechanism, it might be obtained the greater performance to solve the cocktail party problem.

# Bibliography

[1] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, “Two-microphone separation of speech mixtures,” *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 19(3), pp. 475–492, 2008.

[2] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, “Overcomplete blind source separation by combining ICA and binary time-frequency masking,” in *IEEE International workshop on Machine Learning for Signal Processing* (J. L. D. M. S. D. V. Calhoun, T. Adali, ed.), pp. 15–20, sep 2005.

[3] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” 2001.

[4] A. Hyvarinen, “Independent component analysis,” vol. 2, pp. 94–128, 2001.

[5] B. T. Atmaja, “Pemisahan banyak sumber suara mesin dari microphone array dengan metode independent component analysis (ica) untuk deteksi kerusakan.” Tugas Akhir ITS, 2009.

[6] A. Hyvarinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Networks*, vol. 13(4-5), pp. 411–430, 2000.

[7] D. Wang and G. J. Brown, eds., *Computatinal Auditory Scene Analysis: Principles, Algorithms and Application.* John Wiley and Sons.

[8] D. Wang, “Time frequency masking for speech separation and its potential for hearing aid design,” *Trends in Amplification*, vol. 12 (4), pp. 332–353, 2008.

[9] C. Kim, K. Kumar, B. Raj, , and R. M. Stern, “Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain,” *INTERSPEECH*, pp. 2495–2498, 2009.

[10] B. T. Atmaja, T. Usagawa, Y. Chisaki, and D. Arifianto, “On performance of sound separation methods including binaural processors,” in *Student meeting of Acoustic Society of Japan, Kyushu-Chapter*, 2011.

[11] B. T. Atmaja and D. Arifianto, “Machinery fault identification using blind

sound separation and fuzzy system," in *International Fuzzy System Association World Congress - Asian Fuzzy System Association International Conference (IFSA-AFSS)*, pp. FP–003(1–4), June 2011.

[12] B. T. Atmaja and D. Arifianto, "Blind sound separation using frequency-domain and time-domain independent component analysis for machines fault detection," in *Proceeding of The International Conference on Advanced Computing and Information System (ICACSIS)*, pp. 259–263, 2009.

[13] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

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

[15] L. K. Hansen, J. Larsen, and T. Kolenda, "Blind detection of independent dynamic components," *In proc. IEEE ICASSP'2001*, vol. 5, pp. 3197–3200, 2001.

[16] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10(03), pp. 626–634, 1999.

[17] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *ICASSP*, 2011.

[18] M. I. Mendel, *Binaural Model-Based Source Separation and Localization*. PhD thesis, Columbia University, 2010.

[19] H. Nakashima, Y. Chisaki, T. Usagawa, and M. Ebata, "Frequency domain binaural model based on interaural phase and level difference," *Acoustics Science and Technology*, vol. 24(3), pp. 172–178, 2003.

[20] H. Nielsen, "Ucminf - an algorithm for unconstrained, nonlinear optimization," Tech. Rep. IMM-TEC-0019, IMM, Technical University of Denmark, 2001.

[21] M. Tomita, S. Saon, Y. Chisaki, and T. Usagawa, "Quantitative evaluation of segregated signal with frequency domain binaural model," *Acoustics Science and Technology*, vol. 30, pp. 448–451, 2009.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech and Audio Proc.*, 2005. to appear.

# PUBLICATIONS

1. B.T. Atmaja, D. Arifianto. Machinery Fault Identification Using Blind Sound Separation and Fuzzy System. International Fuzzy System Association World Congress – Asian Fuzzy System Society (2011 IFSA-AFSS), Surabaya - Indonesia, pp FP–003(1–4), June 2011.
2. Bagus Tris Atmaja, Yosifumi Chisaki, Tsuyoshi Usagawa, Dhany Arifianto. On Performance of Sound Separation Methods Including Binaural Processors. Student meeting of Acoustic Society of Japan, Kyushu-Chapter, 2011, Oita-Japan, pp. 65-68.
3. B.T Atmaja, D. Arifianto, Y. Chisaki, T. Usagawa. Signal Enhancement by Using Sound Separation Methods Based On Binaural Inputs. Basic Science International Conference 2012, Malang-Indonesia. pp. C1-C6..