On Source Signal Segregation Based On Binaural Inputs

Bagus Tris Atmaja, Dhany Arifianto, Tsuyoshi Usagawa

***Abstract*—Source separation is important thing in such applications like speech recognition, hearing aids and telecommunication. The enhancement task in those application can be done by separating target sound and interference sounds. This sound separation method is interest problem in psychological and computational science motivated by human auditory processing. How the human auditory processing solved this problem is exactly unanswered until new. However, this function of binaural hearing can not be easily transformed to the computational methods. Independent Component Analysis (ICA) is one of existing methods and the fast algorithm is implemented in FastICA. Another approach is by using binary time-frequency masks to obtain the basis signals. This paper evaluate systematically varied methods: ICA, ICA with binary mask, binaural model, FastICA and FastICA with binary mask. Enhanced signals were measured by means of coherence criterion and PESQ score to obtain performance comparison. An implementation of sound segregation was given by machine sound separation task.**

***Index Terms*—Signal enhancement, Source Separation, ICA, Binary Mask, Binaural Inputs**

I. Introduction

Human auditory processing is one of the very smart system on human being which can receive sound and process it using heuristic system to localize, separate, recognize and doing other tasks. One of the capability of binaural hearing using left and right ears is the cocktail party phenomena, in which human being can focus on target sound while listening others sounds. The binaural ears works simultaneously to separate and localize the mixed sounds with the neuro-processing mechanism. The exact method so separate sound sources was studied by psychologist and computer scientist. Some methods were proposed including CASA (Computational Auditory Scene Analysis), ICA (Independent Component Analysis) and binaural model. However, no very robust algorithm was claimed for real data and noises-added data. Some methods are very fast, but it didn't give high score of objective evaluation. Others give fair sound quality after separation, but it has low objective score. This paper evaluate some methods in source separation from the point of view of computational science, modeling and mathematics-statistics. There are five methods to be evaluated in this paper,

1) ICA
2) ICA with binary mask (ICABM)
3) Binaural hearing using phase difference channel weighting (PDCW)
4) FastICA
5) FastICA with binary mask (FastICABM)

All those algorithms was obtained directly from references except FastICA with binary mask. In that method, a combination of FastICA and binary mask was formed and compared to others.

II. Methods

Following five methods which can be used for source segregation task to enhance target signal from binaural inputs are examined in this paper.

A. 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

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

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,

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

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.

B. ICA with Binary Masks

ICA is used with binary time-frequency masking proposed by Wang et al. who were motivated by human auditory phenomenon in which a sound is rendered by louder sound within critical band. The mask $M(n,k)$ in time-frequency domain is expressed as

$$m\left(n,k\right) = \begin{cases} 1 & if\ S_1\left(n,k\right) - S_2\left(n,k\right) > \theta \\ 0 & otherwise \end{cases} \tag{3}$$

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.

C. 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 phase difference in frequency domain to estimate the ITD as described in. 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 each time-frequency bins.

D. FastICA

In this paper, FastICA algorithm introduced by Aapo Hyarinen is used based on [2]. 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 W in this algorithm given by,

$$w^+ = E\left\{xg\left(w^T x\right)\right\} - E\left\{g^{'}\left(w^T x\right)\right\} w \tag{4}$$

$$w = \frac{w^+}{\|w^+\|} \tag{5}$$

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^+.

E. FastICABM

A combination of FastICA described previously was used with binary mask obtained from [3]. The output of FastICA algorithm was processed to estimate binary masks and extract the basis signals from the mixing signals. The diagram of this system was show in Fig. 1. The block diagram was adapted from [4].

In that figure, the two input signals are processed with FastICA algorithm through signal buffer. Because FastICA has ambiguity in output level, the scaling technique was added to normalize the level signal of FastICA's outputs. Then, each of the the two binary masks is applied to the original mixtures in the T-F domain, and by this non- linear processing, some of the sound signals are removed by one of the masks while other speakers are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT [4].

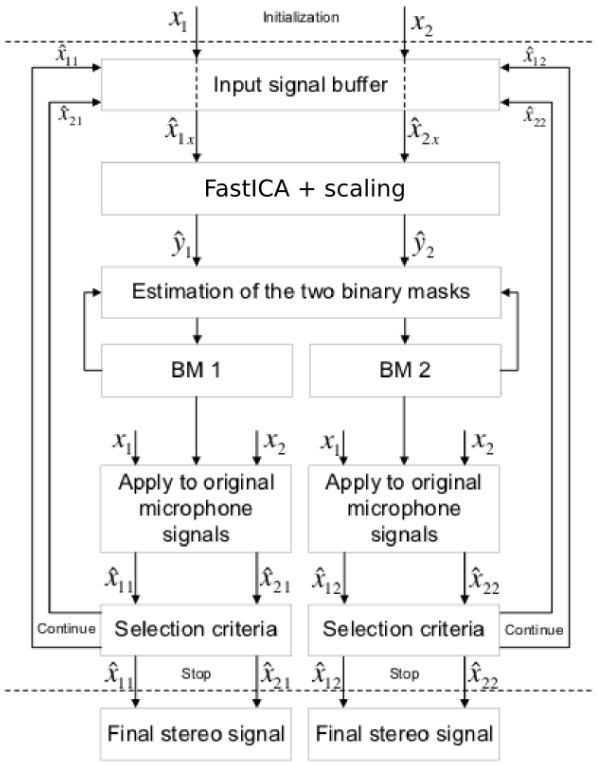

Figure 1. Diagam Block of FastICABM (adapted from [4])

III. Simulation and Experiment

A. Simulation

The simulation to obtain data was performed by convolving HRTF (Head Related Transfer Function) to the source signals according to its azimuth and elevation. The observed signals at left and right are defined as follows,

$$x_L(n) = \sum_i h_l(\theta_i) * s_i + n_l(n) = l_0(n) + l_1(n) + \ldots + n_l(n), \tag{6}$$

and

$$x_R(n) = \sum_i h_r(\theta_i) * s_i + n_r(n) = r_0(n) + r_1(n) + \ldots + n_r(n), \tag{7}$$

where $x_L(n)$ and $x_R(n)$ are observed signal at left and right, $h_r(\theta_i)$ and $h_l(\theta_i)$ are HRTFs from the direction, θ_i and s_i are i-th of source signals. $l_0(n)$ and $r_0(n)$ denote the target signal at left and right, $l_i(n)$ and $r_i(n)$ denote the interference signals, n_l and n_r denote additive noise at left and right. The * sign represent convolution between HRTF and source signals. The observed signals were saved as .wav file with 16000 Hz of sampling frequency and 16 bits PCM. Other sampling frequency are used by resampling method to analyze effect of various sampling frequency.

While equation 6 and 7 show the mathematics formulation of simulation data, Figure 2 shows the simulation and separation process by some methods mentioned above by diagram block for two sound sources. The signals in that figure is

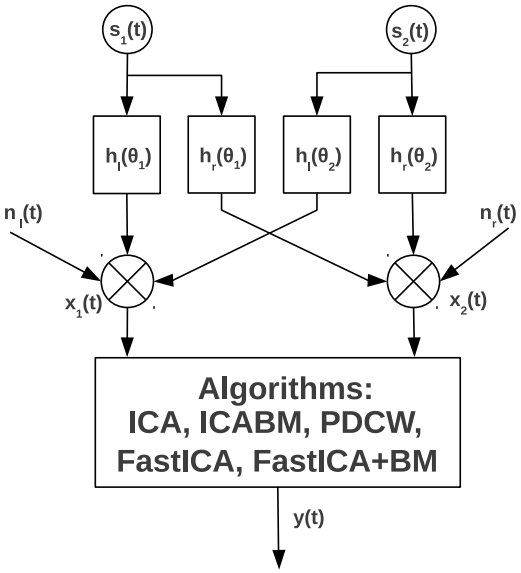

Figure 2. Block diagram of simulation and separation process by some methods

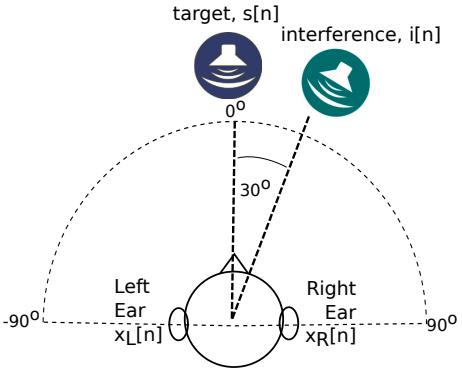

Figure 3. Location of target at $(0°, 0°)$ and interference signals at $(0°, 0°)$

symbolized in time domain. As shown, the separation methods are performed to give enhanced signal $y(t)$ and it is expected to get enhanced signal with the better quality than mixed signal $x_1(t)$ and $x_2(t)$, the signal at left and right channel.

The location of target signal and interference noise for example can be seen in Figure 3. On that illustration, target signal is located on $0°$of elevation and $0°$of azimuth $(0°, 0°)$, while interference signal is in $(0°, 30°)$. On other data, the interference signals are varies from azimuth $-90°$ to $90°$with $10°$ interval.

The target signal is Japanese female speech while the interference signals are Japanese male speech and white noise signal. The enhanced signal represented female speech after separation task and it was compared to the clean female speech by means of coherence criterion based on [5] and PESQ score based on [6].

B. Experiment

Experiment is performed under the setup shown in Fig. 4. Two loudspeakers are located in the direction of 0^o and -45^o of a head and torso simulator (HATS) as shown in Fig. 4. The target signal is female speech where white noise is used

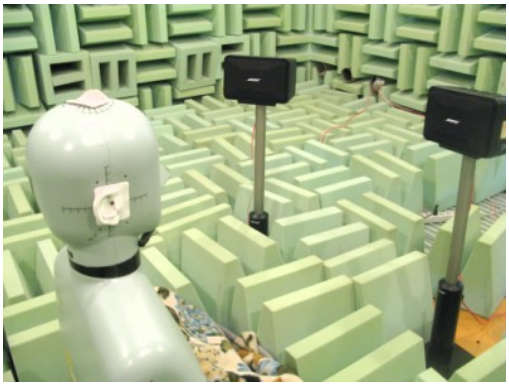

Figure 4. Screenshot of experiment in anechoic room

for interference. The level of interference is set to the same 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.

IV. Result and Discussion

The separation task and its result analysis can be divided into five groups: result from simulation and experiment data, types of interference, effects of various SIR, effect of various SNR and effect of various sampling frequency. Those various condition will be analyzed and discussed one by one.

Result of Simulation and 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 [3]. 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. IV.

Types of Interference

The second task was performed with two different types of interferences. 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

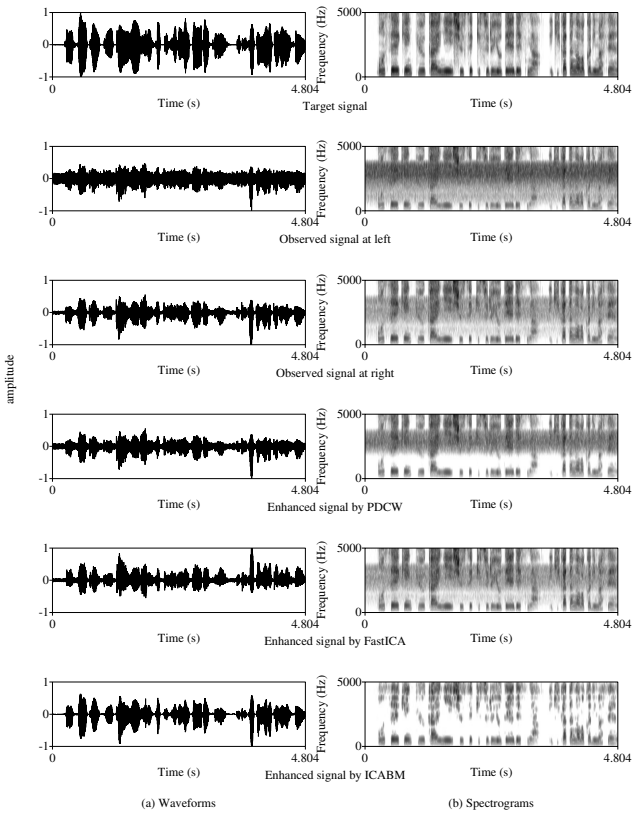

Figure 5. Signals and its spectrogram in experiment

Table I
COMPARISON OF AVERAGED COHERENCE SCORE BETWEEN SIMULATION AND EXPERIMENT

Method	Simulation	Experiment
$PDCW$	0.542	0.283
$FastICA$	0.669	0.351
$ICABM$	0.539	0.277

signals from male speech interference and separation of target signal from male speech and with noise interferences. The results of those condition can be shown on Table II, III and IV respectively.

In Table II, result of separation task from target signal corrupted with white noise interference was shown. The SIR (target vs interference) was set to 20 dB while SNR (target vs additive noise) is 10 dB. Noise was included to simulation to close the real environment although some research didn't include noise and assume ideal condition. From this condition, the highest coherence score was obtained by FastICABM and the PESQ scores are almost similar of five methods.

Table II
SEPARATION TASK OF TARGET SIGNAL AND WHITE NOISE INTERFERENCE (SIR= 20 DB, SNR= 10 DB)

Methods	Coherence	PESQ
ICA	0.724	1.940
ICABM	0.683	1.946
PDCW	0.578	1.906
FastICA	0.724	1.938
FastICABM	0.792	1.905

Table III
SEPARATION TASK OF TARGET SIGNAL AND MALE INTERFERENCE (SIR= 20 DB AND 4 DB, SNR= 10 DB)

Methods	Coherence	PESQ
ICA	0.735	2.078
ICABM	0.715	2.495
PDCW	0.554	1.562
FastICA	0.734	2.075
FastICABM	0.715	2.460

Table IV
SEPARATION TASK OF TARGET SIGNAL WITH MALE AND WHITE NOISE INTERFERENCE (SIR= 4 DB AND 20 DB, SNR= 10 DB)

Methods	Coherence	PESQ
ICA	0.724	1.748
ICABM	0.683	2.023
PDCW	0.579	1.332
FastICA	0.724	1.749
FastICABM	0.720	2.010

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. Four methods have PESQ score above 2.0. Only PDCW has 1.562 of PESQ score. In this paper, PDCW method has the lowest score in coherence criterion and PESQ score. The method previously used 4 cm of microphone distance, 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 occurred in PDCW and it reduced the separation quality of PDCW method.

The last condition of evaluating different types of interference was performed by corrupting female speeh target with male and white noise simultaneously. The male speech is located in $(0°, -30°)$ and white noise interference is in $(0°, -30°)$. The result from that condition is figured out in Table IV. The value of SIR and SNR are set to be same to the previous value. In this case, ICABM has the highest score of PESQ white the highest score in coherence was obtained ICA and FastICA which has the same result.

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.

Effects of Various SIR

The third task 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 is chosen as interference signal located in $(0°, 30°)$

Table V
COMPARISON OF COHERENCE CRITERION FROM SEPARATION TASK IN VARIOUS SIR

Methods	SIR				
	-20 dB	-10 dB	0 dB	10 dB	20 dB
ICA	0.598	0.598	0.597	0.633	0.633
ICABM	0.603	0.608	0.394	0.325	0.325
PDCW	0.513	0.500	0.409	0.213	0.213
FastICA	0.598	0.598	0.597	0.632	0.632
FastICABM	0.631	0.609	0.315	0.418	0.471

Table VI
COMPARISON OF PESQ SCORE FROM SEPARATION TASK IN VARIOUS SIR

Methods	SIR				
	-20 dB	-10 dB	0 dB	10 dB	20 dB
ICA	1.180	1.180	1.184	1.378	1.378
ICABM	1.185	2.077	1.548	0.692	0.692
PDCW	1.169	1.167	1.190	0.991	0.991
FastICA	1.180	1.180	1.184	1.379	1.379
FastICABM	1.268	2.112	1.282	0.935	1.268

while target signal still in $(0°, 0°)$. The separation result by means of coherence criterion can be shown in Table 4 which shows all result from five methods in various SIR. As can be seen in that figure, on -20 dB and -10 dB of SIR, result of FastICABM has the highest coherence score. However, when SIR between target signal and interference signal changed to 0 dB, 10 dB and 20 dB, conventional ICA method obtain the highest score of coherence. Table 4 also shows that there are almost no different of the result between -20 dB and -10 dB and also between 10 dB and 20 dB especially for ICA, PDCW and FastICA method. This result need to be check again with human auditory behavior and experiment data or it might be caused by sensitivity of algorithm in which the algorithm can not differ SIR from -20 dB and -10 dB also 10 dB and 10 dB.

The next table, Table 5, shows result of separation task in different in various SIR in PESQ score. Again, at the -20 dB and -10 dB of SIR, FastICABM obtain the highest score, while at 0 dB highest PESQ score was obtained by ICABM and at 10 dB and 20 dB, conventional ICA reach the highest value among others. The result of ICABM close to perceptual evaluation by listening the enhanced sound. So, PESQ score is more suitable when the purpose of sound separation is designed for perceptual application such as in hearing aids and telecommunication.

Effects of Various SNR

In the task of sound sources segregation problem using various value of SIR, the value of SNR was fixed to 20 dB. This research takes into additive noises assumed from background noises or other noises like from hardware which usually not be included in other research like in [2], [4], [3]. This noises can not be avoided in real environment, so it must be included to close the real source separation problem.

Six different SNR values were evaluated from 0 dB to 25 dB with interval of 5 dB. The result show increasing of objective evaluation score for both coherence and PESQ score. As shown in Figure 3, at 0 dB of SNR, the averaged coherence criterion is 0.445 while the PESQ score at that condition is

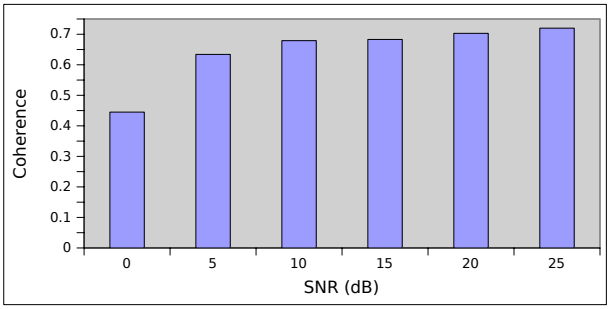

Figure 6. Comparison chart of coherence criterion in various dB SNR (target vs additive noise)

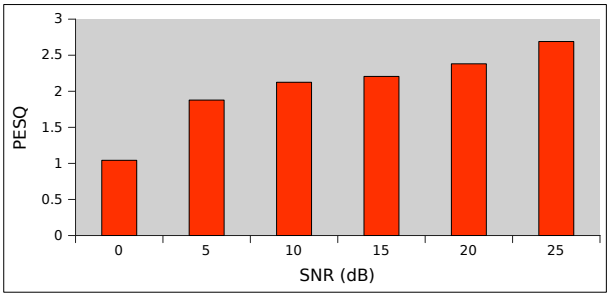

Figure 7. Comparison chart of PESQ score in various dB SNR (target vs additive noise)

1.043. The highest sampling frequency was achieved at the highest dB SNR i.e 20 dB with averaged coherence value of 0.72 and PESQ score of 2.69. In the opposite of previous task in section IV this task was performed with fixed value of signal to interference ratio (SIR) i.e 4 dB between female speech and male speeh interference. This result show the additive noises effect separation result which in other researches are assumed ideally without noises to simply the separation task. In this task, FastICABM algorithm was used to evaluate separation result of different SNR and the obtained PESQ score is the highest among all data and task presented in this paper.

Effects of Various Sampling Frequency

The last task of separation problem in this paper was performed by evaluating different sampling frequency to the sound sources and HRTF. The standard sound files from simulation data recorded in 16000 Hz are downsampled and upsampled to be 8000 Hz, 22050 Hz, 44100 Hz and and 48000 Hz. The HRTF which has 44100 Hz of sampling frequency is also downsampled and upsampled to those kinds of sampling frequency. The different sampling frequency data then was used as input of ICA algorithm. The objective evaluation for this task is only given by coherence criterion because PESQ was designed for 8000 Hz and 16000 Hz of sampling frequency only. Other algorithm like PDCW also cannot processed data above 16000 of 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. However, using 16000 Hz of sampling frequency in real time processing such as in telecommunication needs more effort.

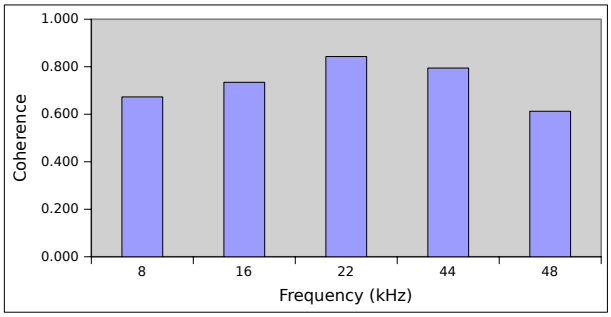

Figure 8. Comparison chart of coherence criterion in various sampling frequency

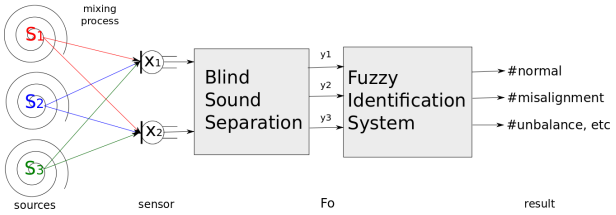

Figure 9. Proposed machine fault identification system

The calculation of cost function should be considered for real application purposes. At the 22050 Hz of sampling frequency, the highest coherence criterion was 0.84 which is the highest value among all data presented in this paper.

Implementation in Machine Sound Separation

In this last section, an implementation of sound separation method to separate machine sounds was proposed as diagnostic tools. Microphone array was used as sensor and electrical motor represented machine in small scalel. Four common conditions in machine were evaluated i.e normal, unbalance, misalignment and bearing fault. Microphone array will record the signal sound and source separation algorithm will separate the mixed signals to estimate each machine sounds. As control data, the machine sound was recorded individually in each condition. For idetification step after separation process, fuzzy logic was proposed along with BSS/ICA methods as shown in Figure 9. The if-then rules then propose as the following,

- If F_0 =32.53-69.67 and F_1 =1757-1789 and F_2 =2618-2696 then the machine is NORMAL
- If F_0 =27.62-58.28 and F_1 =995-1005 and F_2 =1732-1808 then the machine is UNBALANCE
- If F_0 =21.31-67.62 and F_1 =1764-1782 and F_2 =2501-2697 then the machine is MISALIGNMENT
- If F_0 =22.17-117.24 and F_1 =200.50-285.00 and F_2 = 303.90 - 379.80 then the machine is BEARING FAULT

The rules show that we used three frequency as input for fuzzy system. The range value of F_1, F_2 and F_3 are obtained from experiment. The implementation of the fuzzy system is to identify whether machine is normal, unbalance, misalignment or bearing fault. Because instantaneous frequency approach was used to estimate fundamental frequency, so the frequency is changed with respect to time.

V. Conclusions

This paper evaluated some task and various condition of source separation problems for signal enhancement purpose. The first task evaluated the separation result from simulation data and experiment data. In that task, FastICA has the highest coherence criterion but ICABM gives fair sound quality although it is degraded. The different types of interferences also 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. In 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 linear result was obtain when increasing SNR to mixed sound. The higher dB SNR between target signal and additive noise, the better separation result obtained. Noises actually can't be avoided and it affected the separation result as presented in this paper. Finally, the use of different sampling frequency gives the highest score of coherence criteria at 22050 Hz. To choose objective 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 psychoacoustic model and cognitive model. Finally, an example of implentation of sound source segregation was given by machine source separation. The task is to separate machine sound sources from mixed sounds. As identification sytem, fuzzy logic was proposed based on fundamental frequency obtained from separated sounds.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] I.-T. R. 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.