A consideration on time-frequency masking methods for speech separation

Ning DING†, Takayuki SHIMADA†, Masahshi YOSHIDA†, Junya ONO†,
Wlodzimierz KASPRZAK‡, Nozomu HAMADA†

†Signal Processing Lab., School of Integrated Design Engineering, Keio University
‡Institute of Control and Computation Engineering, Warsaw University of Technology

ABSTRACT Time-Frequency Masking methods, primary known as DUET [2] and SAFIA [3], are effective scheme for blind speech separation problem. Based on an investigation of conventional delay-histogram and the time-frequency masking method in terms of estimated delay accuracy, two novel approaches for clustering process are proposed. In particular, the proposed methods tend to improve relatively large amount of delay estimation error using STFT phase difference in lower frequency band. A novel clustering idea is the use of frequency vs. phase difference dot plot in a frame by frame manner. The other one is to use filter bank and fractional delay operation. These approaches are proved to be effective through several experiments.

1 Introduction

To realize speech-based human-machine interfaces, such as high quality hand free communication, speech separation has been considered to have a very important role. Blind Source Separation (BSS) is an approach for estimating source signals by using only the mixed signals observed at each input channel. The estimation is performed blindly, without possessing information on each source, such as its location and active time. Typical examples of such source signals include mixtures of simultaneous speech signals that have been picked up by several microphones [1].

Many methods have been proposed for BSS problem, such as Independent Component Analysis (ICA) and Time-Frequency (T-F) masking. ICA is relied on statistical independence of the speech signal if signals are mixed instantaneously. However, it is difficult for ICA to solve the underdetermined case in which source number is greater than microphone number.

Time-Frequency masking method, known as DUET[2] as its primary representative, transfers the signal from time domain to time-frequency domain by Short Time Fourier Transform (STFT). It is based on the assumption called "W-Disjoint Orthogonality (WDO)" [2], which means though the observed signal is mixture of several sources, most part of the time-frequency cells contain at most one of the source signals' component.

Some assumptions are also proposed in the method known as SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones) [3] for separation speech signals. In addition to these methods, several novel schemes are developed recently, such as DEMIX [4], and TIFROM (TIme-Frequency Ratio Of Mixtures) [5].

This paper mainly focuses on the time-frequency masking based on delay histogram and provides some improvements to it. There are some drawbacks in the delay histogram approach. One of them is the phase misestimate, especially in low frequency band.

The paper is organized as follows. In Section 2, the BSS problem is briefly reviewed. In Section 3, we introduce the frequency vs. phase difference scatter plot of time-frequency cell in a frame-by-frame mannar and it is used for clustering in Section 4. Another clustering approach combining filterbank and fractional delay operation is proposed in Section 5. Some experiments are taken to verify our method. Section 6 is the conclusion.

2 BSS problem

In discrete time domain, suppose that sources s_1 , $\cdots$,s_N are convolutively mixed and observed at M sensors

$$x_j(\tau) = \sum_{k=1}^{N} \sum_{l} h_{jk}(l) s_k(\tau - l) \quad j = 1, \ldots, M \qquad (1)$$

where $h_{jk}(l)$ represents the impulse response from source k to sensor j, N is the number of sources, and M is the number of sensors.

Signal transformation to time-frequency domain is like this: time domain signals $x_j(\tau)$ sampled at frequency f_s are converted into frequency domain time-series signals $X_j(t, f)$ with a L point STFT:

$$X_j(t,f) = \sum_{r=-L/2}^{L/2-1} x_j(r+tS)win(r)e^{-i2\pi fr} \quad (2)$$

where $win(r)$ is a window, S is the window shift size, t is the integer time frame index, and f is the integer $(0 \sim \frac{L}{2})$ frequency bin.

Time-frequency approach utilizes instantaneous mixtures at each time frame t and frequency bin f:

$$X_j(t,f) \approx \sum_{k=1}^{N} H_{jk}(f)S_k(t,f) \quad (3)$$

where $H_{jk}(f)$ is the frequency response, and $S_k(t,f)$ is a frequency domain time-series sourse signal.

In time-frequency domain, signals have the property of sparceness. In mathematical form, it is described as

$$S_1\ (t,f) \cdot S_2\ (t,f) \approx 0 \quad \forall (t,f) \quad (4)$$

3 Cell's features and clustering method

Separation is realized by cell clustering algorithm and inverse STFT. The binary mask approach depends strongly on the clustering of the feature, so the selection of an appropriate feature is essential to this approach.

3.1 Features

Most common method utilizes the delay calculated from the phase difference between observations as their cell's features. In DUET [2], they define a power weighted two-dimensional (2-D) histogram constructed from the ratio of the time-frequency representations of the mixtures, which is shown to have one peak for each source with peak location corresponding to the relative attenuation and delay mixing parameters. This histogram is used to cluster time-frequency cells or equivalently to generate separation masks. In SAFIA[3], for each component, differences in the amplitude and phase between channels are calculated as in DUET. These features are used to select frequency components of the signal that comes from the desired direction and to reconstruct these components as the desired source signal. In MENUET [6], their method bases on the normalization and clustering of the level ratios and phase differences between multiple observations. In HS [9], they propose harmonic structure as clustering feature. To estimate the harmonic structure, the proposed method estimates the fundamental frequency using the initial separation, the candidate detection and fundamental selection. Finally, the harmonic structure mask is combined with DOA mask.

In this paper, we focus particularly on a situation where the number of sources $N = 2$, and the number of sensors $M = 2$. Our approach is using time-frequency masking methods based on DOA. In the following, we will introduce two basic classification processes: delay calculation, and binary mask generation.

3.2 Delay calculation

Anechoic mixing process can be expressed as

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ e^{-j\frac{2\pi f\delta_1}{L}} & e^{-j\frac{2\pi f\delta_2}{L}} \end{bmatrix} \begin{bmatrix} S_1\ (t,f) \\ S_2\ (t,f) \end{bmatrix} \quad (5)$$

δ_i (i =1,2) is the delay between two microphones, and L is the number of STFT points. Assuming microphone 1 is the reference point, under the condition of WDO, it can be simplified to

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ e^{-j\frac{2\pi f\delta_i}{L}} \end{bmatrix} S_i\ (t,f) \quad (6)$$

The delay δ_i is obtained using phase correlation function[7]:

$$\delta(t,f) = \frac{L}{2\pi f}\phi(t,f) \quad (7)$$

where $\phi(t,f)$ is the phase difference,

$$\phi(t,f) = \angle X_1(t,f) - \angle X_2(t,f) \quad (8)$$

3.3 Delay histogram-based separation

Since speech signal has sparsity property against both time and frequency, to reconstruct the original signals, time-frequency cells must be clustered into two groups. The delay between observed signals can be an effective feature. By using the estimated delay and its histogram, in which two peaks, δ_1 and δ_2 are attained, corresponding to two sources. Though the delay data $\delta(t,f)$ are spread, the peaks can approximately estimate the direction of sources. Then binary masks are generated by

$$M_1(t,f) = \begin{cases} 1 & if\ |\delta(t,f) - \delta_1| < |\delta(t,f) - \delta_2| \\ 0 & otherwise \end{cases} \quad (9)$$

$$M_2(t,f) = \begin{cases} 1 & if\ |\delta(t,f) - \delta_1| > |\delta(t,f) - \delta_2| \\ 0 & otherwise \end{cases} \quad (10)$$

Therefore, the speech mixture signal could be separated by binary masks $M_i(t,f)$, and the separated signals $\hat{S}_i(t,f)$ is given by the following masking.

$$\hat{S}_i(t,f) = M_i(t,f)X_j(t,f) \quad (11)$$

Finally, by Inverse Short Time Fourier Transform (ISTFT), the separated signals obtained in time domain.

3.4 Phase difference scatter plot

Although time-frequency masking based on delay is a good method for BSS problem, in real circumstance, there are many errors caused by phase difference estimation. For example, the caluculated delay time derived from phase difference between two microphone signals should be less than d/c, where d is the distance between microphones, c is the sound velocity, but the estimated delay violates this restriction. Fig.1(a) shows an examle of the scattered diagram on phase difference versus frequency $\{f, \phi(t,f)\}$ for several frames. Each dot corresponds to each T-F cell. Fig.1(b) indicates the relative average error of phase difference is nearly equal. This and equation (7) cause the larger amount of error in delay for the lower frequency as shown in Fig.2. In conventional method, the clustering is given by drowing the separation line at the middle of two histogram peaks.

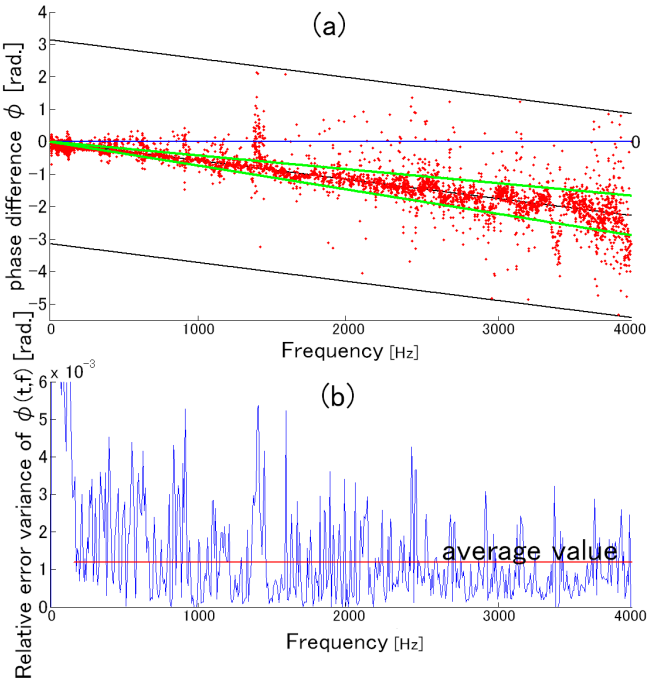

Fig. 1: Frequency vs. phase difference diagram (a)Scattered dot $\{f, \phi(t,f)\}$ plot. (b)Relative error variance of $\phi(t,f)$ at each f. Real environment data (source angle=50°, $\bigcup t$ = 92frames)

Due to nearly uniform relative error variation on $\phi(t,f)$, clustering by the use of delay will be very difficult in the lower frequency band. Namely, in lower frequency band, the distributions of $\{f, \phi(t,f)\}$ dots corresponding to two sources would be overlaped, it is not possible to separate by means of dot position.

One way to cope with the misestimation in low frequency band is deleting these lower frequency components when generating binary mask. If we set the cut off frequency very low, the separated signal will still contain the component of misestimate. On the other hand, if we set the cut off frequency very high, it will affect the tone quality of separated signals. The selection criterion of cut off frequency is keeping the tone quality of separated signals, at the same time, depressing misestimate components as much as possible. As demonstrated in reference [9]and by a lot of experiments, the cut off frequency is set to 400Hz.

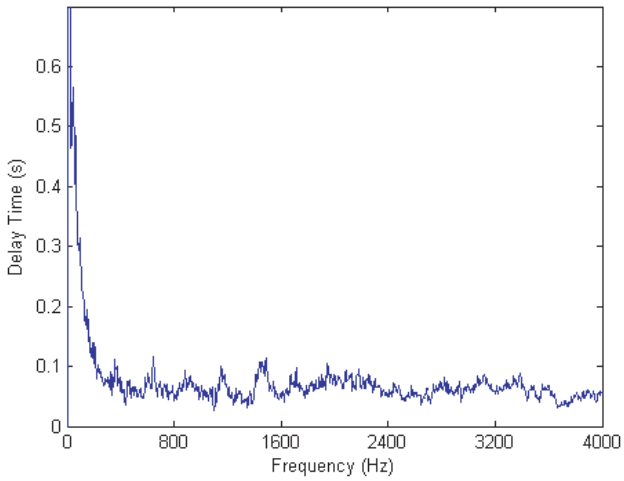

Fig. 2: Avaraged delay estimation at each frequency bin

4 Proposed method

The investigation in previous sections teaches us that the cell clustering at lower frequency bin (we denote $f \in L_\omega$ such as, $\{f | f < 400Hz\}$) using delay is inherently erroneous and its misestimation finally causes poor separation performances. To solve this problem, this study focuses on the use of individual scattered map at each frame, namely $\{f, \phi(t,f); t = t_n\}$ and proposes novel clustering at lower frequency cell components. The proposed idea is to utilize relatively reliable middle and higher frequency component's attribution in order to determine that of lower frequency cells. Considerable strategies are (i) single source active (SSA) frame detection, and (ii) harmonic structure detection. Here, we focus solely on the former case.

Thus, the task we should introduce is the way to check whether a given frame is the SSA one or not. In addition, for detected SSA frame, the determination on

which source is active at that frame should be done.

4.1 SSA detection

As we can see from the $\{f, \phi(t,f)\}$ dot plot in Fig.1(a), whether the given frame is at SSA state or not would be reflected as a scattering feature along a constant gradient line, i.e. a line with a specific delay. If the given frame is at simultaneous utterance interval, the scattered dots tend to distribute along two different lines $\phi(t,f) = (p_1, p_2 : real\ number)$. Each of these corresponds to each source direction.

The principal component analysis (PCA) is applied to check SSA feature. The adopted criterion is that one of the eigenvalues (the second eigenvalue) of 2-dimentional data $\{f, \phi(t,f)\}_{t=t_n}$ is sufficiently smaller than another (the first) eigenvalue.

The process determining the attribution of lower frequency components$\{X_i(t,f), f \in L_\omega\}$ are given as follows. The two source directions are obtained without using lower frequency bin ($f \in L_\omega$) beforehand, and these directions give corresponding delays δ_1, δ_2.
(i)Apply the PCA to the data $\{f, \phi(t,f)\}_{t=t_n}$ $f \in L_\omega$
(ii)For the normalized eigenvalues $(1, \epsilon_2)$(the normalization means that the larger eigenvalue ϵ_1 is set to unity), the following criterion is applied to check the SSA at frame t_n.

$$\begin{cases} |\epsilon_2| < Th_1 & t_n \text{ frame is SSA} \rightarrow \text{go to (iii)} \\ |\epsilon_2| > Th_1 & t_n \text{ frame is non SSA} \end{cases} \quad (12)$$

(iii)For the SSA frame, the gradient a_1 of the first principal axes is used to determine the attributes of cell components in the lower frequency band. The cell's attributes in L_ω is represented by

$$\arg\min_n |a_1 - p_n| \quad \text{for} \quad |a_1 - p_n| < Th_2 \quad (n = 1, 2) \quad (13)$$

where p_n are the gradients of two linear phase difference $\frac{2\pi f}{L f_s}\delta_n$

Above clustering procedure is solely applied to $f \in L_\omega$ in our approach. The rest cell attributions are allocated by the conventional delay histogram scheme. The combination of the proposed and previous clustering procedures generates binary separation masks M_1 and M_2, and completes the source separation.

4.2 Experimental condition

Some experiments are performed in a conference room to certify our methods. The geometrical parameters are shown in Fig.3, and other parameters are shown in Table 1. We use nine kinds of ASJ continuous speech corpus [10] for research as the source signals. One source is located at the broadside (0 degree) and the other sources are located from 10 degrees to 90 degrees at every 10 degrees.

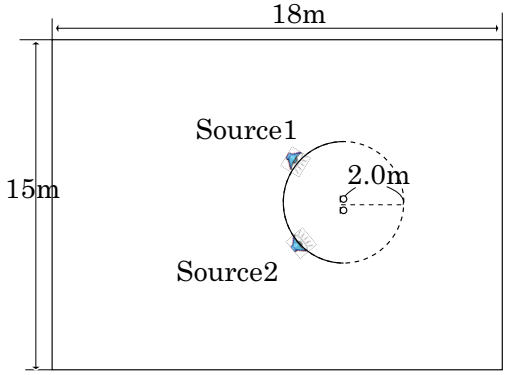

Fig. 3: Geometrical parameters

Table 1: Experiment parameters

Sampling Frequency	8000HZ
Microphone Distance	40mm
Sound Velocity	340m/s
Window	Hamming
STFT Frame Length	1024
Frame Overlap	512

4.3 Experimental results

The parameters Th_1 and Th_2 of separation process are set to $Th_1 = 0.065$, $Th_2 = 0.005$. As the value of performance evaluation, we use WDO (measure of W-disjoint orthogonality)[2]. It is computed from two other criteria PSR (the Preserved-Signal Ratio) and SIR (the Signal-to-Interference Ratio) defined as,

$$WDO = \frac{||M(t,f)S_d(t,f)||^2 - ||M(t,f)S_i(t,f)||^2}{||S_d(t,f)||^2} = PSR - \frac{PSR}{SIR} \quad (14)$$

$$PSR = \frac{||M(t,f)S_d(t,f)||^2}{||S_d(t,f)||^2} \quad (15)$$

$$SIR = \frac{||M(t,f)S_d(t,f)||^2}{||M(t,f)S_i(t,f)||^2} \quad (16)$$

Where $S_d(t,f)$ is the desired signal, $M(t,f)$ is the binary mask, and $S_i(t,f)$ is the interfering signal. Usually, $0 \leq WDO \leq 1$. For ideal separation, it gets $WDO = 1$.

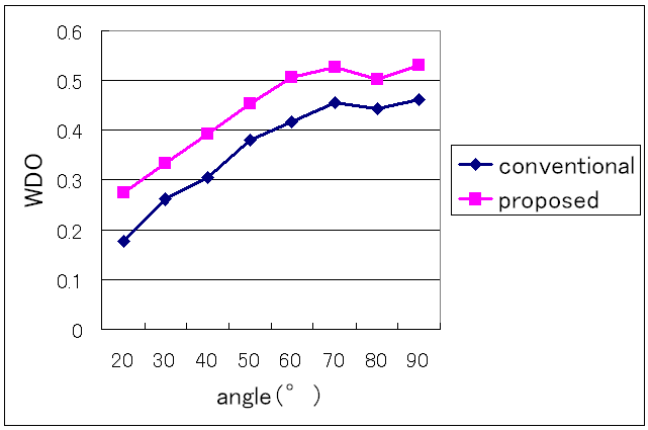

Fig. 4: experomental results

Fig.4 shows the results of the proposed and conventional delay-histgram methods. The proposed method gives nearly 0.05 higher WDO than that of conventional.

It is obvious that the advantage of the proposed method so much depend on the rate of SSA state time interval within whole processed time interval.

5 Cell classification using subsample delay operator and filter bank

5.1 Clustering process

In the conventional cell classification, the distances between the STFT delay $\delta(t,f)$ at each T-F cell and two peak values of delay histogram, denoted δ_1 and δ_2 as in 3.3, are computed as shown in eqs.(9) and (10).

The second proposed method in this study is applied after obtaining δ_1, δ_2 as mentioned above. The basic idea is also based on the sparcity of speech T-F spectrum. But, instead of STFT phase difference, filter bank and delay operators are utilized.

The procedure is described in detail as below.

(i) Two delayed signals of a microphone besides of the reference, $x_2(\tau)$ by amount of δ_1 and δ_2 are generated. As defined by eq. (7), $\delta(t,f)$ is always less than sampling interval $\frac{1}{f_s}$, namely fractional or subsample delay. In this paper, the following fractional delay operation is applied.

$$x_2^i(\tau) = \text{delayed signals of } x_2 \text{ by } \delta_i \cong \sum_{k=-K}^{K} \mathrm{sinc}(kT_s - \delta_i) x_2(\tau - k) \ (i = 1, 2) \quad (17)$$

where, Ts is the sampling interval.

(ii) Bandpass components of $x_i^2(\tau)$ are derived by means of equal bandwidth filter bank. In this system, each channel of the bandpass filter is correspond to the frequency bin f $(0 \sim \frac{L}{2})$, and $150th$ order linear-phase FIR filters with 7.8Hz (-3dB) bandwidth are used.The outputs of $(\frac{L}{2}+1)$ channels filter bank are denoted by $y_2^i(\tau, f)$.

(iii) Multiplying a window function $win(t)$, to filter outputs gives

$$z_2^i(\tau, f) = y_2^i(\tau + tS, f) win(\tau) \quad (18)$$

(iv) The filter band outputs for input $x1(\tau)$ are denoted by $y1(\tau, f)$, and windowed signals are given by

$$z_1(\tau) = y_1(\tau + tS, f) win(\tau) \quad (19)$$

The following criteria are introduced for clustering process using cross correlations.

[Clustering rule]

For a cell component (t,f), if

$Corr.\{z_1(\tau,f), z_2^1(\tau,f)\} > Corr.\{z_1(\tau,f), z_2^2(\tau,f)\}$

, where $Corr.\{\}$ means the normalized cross correlation function, is satisfied, the cell component (t,f) should be allocated to source 1. Otherwise, cell component is added to source 2.

5.2 Experimental results

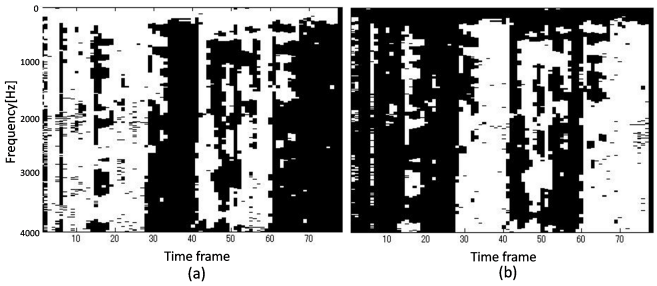

Fig. 5: T-F mask

The proposed method is applied to speech sound captured in real environment in the same conditions as listed in Table 1. Fig.5 (a) , (b) show a derived T-F mask, and WDO values for each source direction respectively. From these results, even the proposed method still does not give sufficient improvement on separation. However, when we evaluate WDO values solely within higher $(f > 400\text{Hz})$ frequency band, the proposed method performs well.

6 Conclusion

This paper investigate the Time-Frequency masking method based on DOA, analyze the influence of phase

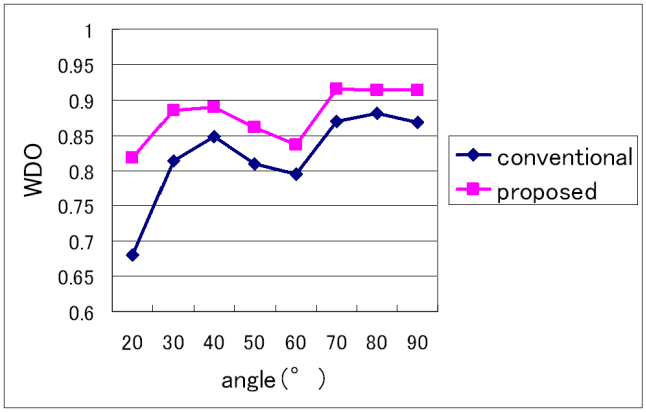

Fig. 6: experimental results

difference misestimate, and propose methods to improve the conventional algorithm. The first method uses the phase difference features with respect to frequency given by scattered plot, and frame by frame operation provides effective clustering when the frame is at single source active state. The second method utilizes filter bank and fractional delay processing for clustering. The experiments show that the proposed methods can improve the separation.

Reference

[1] S. Makino, H. Sawada, R.Mukai, and S. Araki、"Blind source separation of convolutive mixture of speech in frequency domain" *IEICE Trans, Fundamentals*, Vol. 88, No. 7, pp.1830-1847, 2004

[2] O. Yilmaz and S. Rickard、"Blind Separation of Speech Mixtures via Time-Frequency Masking" *IEEE Trans. On signal processing*, Vol. 52, No. 7, pp.1830-1847, 2004

[3] M. Aoki, M. Okamoto S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda、"Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones" *Acoust. Sci. & Tech*, Vol. 22, No. 2, pp.149-157, 2001.

[4] S. Arberet, R. Gribonval, and F. Bimbot、"A robust method to count and locate audio sources in a stereophonic linear anechoic mixture" *ICASSP 2007*, Vol. 3, April 2007, pp. 745-748.

[5] Frederic Abrard, and Y. Deville、"A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources" *Signal Processing*, Vol. 85, pp.1389-1403, 2005.

[6] A. Araki, H. Sawada, R. Mukai, S. Makino、"Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors" *Signal Processing*, Vol. 87, pp.1833-1847, 2007.

[7] A. Averbuch, Y. Keller、"FFT based image registration" *ICASSP'02*, pp.3608-3611, 2002, Orlando, Florida.

[8] Y. Takenouchi and N. Hamada、"Time-Frequency Masking for BSS Problem using Equilateral Triangular Microphone Array" *IEEE Proceedings*, pp.185-188, Hong Kong, 2004

[9] H. Ouchi and N. Hamada、"Separation of Speech Mixture by Time-Frequency Masking Utilizing Sound Harmonics" *Journal of Signal Processing*, Vol. 13, No. 4, pp.331-334, July, 2009.

[10] T. Kobayasi, S. Itahashi, S. Hayamizu and T. Takezawa、"ASJ continuous speech corpus for research [in Japanese]" *The Journal of the Acoustical Society of Japan*, Vol. 48(12), pp.888-893, December, 1992.