BLIND SOURCE SEPARATION BASED ON ACOUSTIC PRESSURE DISTRIBUTION AND NORMALIZED RELATIVE PHASE USING DODECAHEDRAL MICROPHONE ARRAY

Motoki OGASAWARA, Takanori NISHINO, and Kazuya TAKEDA

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
phone: +(81)52-789-4432, fax: +(81)52-789-3172,
email: {ogasawara,takeda}@sp.m.is.nagoya-u.ac.jp, nishino@esi.nagoya-u.ac.jp
web:http://www.sp.m.is.nagoya-u.ac.jp/˜ogasawa/

ABSTRACT

We developed a small dodecahedral microphone array device and propose a sound source separation method based on frequency-domain independent component analysis with the developed device. The developed device's diameter is 8 cm and the intervals among each face are 36°. Microphones can be installed on ten faces except the top and bottom faces, and 16 holes exist on each face. Our proposed method solves the permutation problem, which is frequency-domain independent component analysis's difficult problem, with acoustic pressure distribution that was observed in the device's faces and the normalized relative phases at each microphone in the high and low frequency ranges, respectively. In our experiments, three kinds of mixture signals were used. The separation performances were evaluated by the signal-to-interference rate improvement score and compared with the conventional method and the ideal condition. The results indicate that the proposed method using the developed device is effective.

1. INTRODUCTION

An extraction of the sound source and an estimation of the sound-source direction, which are termed "encoding acoustic fields," are important techniques for many applications, for example, high-realistic communication systems, speech recognition systems, tele-conference systems, and so on. A free-viewpoint TV (FTV) system [1] is one high-realistic communication system that can generate images at a desired viewpoint. For its audio part, a selective listening point (SLP) audio system was proposed that can provide a sound field at an arbitrary selected listening point [2]. SLP audio system is based on an extraction of sound source signal and a stereophonic technology to reproduce a sound field. This system can work on the condition that the number and locations of the sound sources are unknown. As another example, a real-time multimodal system for analyzing group meetings [3] has been proposed that can estimate speaker diarization, for example, "who is speaking and when" and "who is looking at whom" by audio and image signals. Since users can emphasize and listen to selected speech, this system is also considered an acoustic field reproduction scheme. These systems are composed of a source separation method and an estimation method of the sound-source direction and performance reflects the accuracy of the encoded acoustic field.

Frequency-domain independent component analysis (FD-ICA) [4] is usually used for source separation; however, it has a difficult problem called the permutation problem, which many methods have been proposed to solve [2, 5, 6]. A method for using the separated signals themselves has also been proposed [5]. This method supposes that all different frequency components from the same signal are under the influence of a similar modulation in amplitude; however, this assumption is not always correct. In another method that uses the spatial information of the acoustic field [6], estimating the sound-source directions is important. The arrangement of microphones is crucial due to employing time delays among them, and separable source signals are restricted by the location and arrangement of the microphone array. To handle this problem, the SLP audio used many microphones and arrays that surrounded the acoustic field and grouped geometrically similar frequency components together [2]. This method was effective, however, such alignment of microphones and microphone arrays is not practicable. Since SLP audio is one part of the FTV, microphones must not obstruct the view. Therefore, a new microphone array system must be developed to achieve easy alignment and an unobtrusive shape.

In this paper, we develop a novel sound receiving system and propose a method to solve the permutation problem in FD-ICA. A small dodecahedral microphone array device was developed to achieve more robust separation when there are many sound sources. This device that is approximated a sphere can deal with sound sources located in any place. Moreover, this device can be installed in many microphones and is easy to set up. In our method, the permutation problem was solved using acoustic differences that were observed on the developed device's faces. The performances of the sound source separation were evaluated objectively and compared with the conventional method that was proposed in previous research [7].

2. DODECAHEDRAL MICROPHONE ARRAY

A dodecahedron, which resembles a sphere, is usually used for acoustic measurement systems such as loudspeaker systems and microphone arrays. Figure 1 shows our developed dodecahedral microphone array device. This device is designed by computer aided design (CAD) and modeled in acrylonitrile-butadiene-styrene (ABS) resin by a 3D printer (STRATASYS Dimension). The developed device's diameter is 8 cm and the intervals among each face are 36°. Microphones can be installed on ten faces except for the top and bottom faces, and 16 holes exist on each face. The distance between the center of each hole on the same face is 7

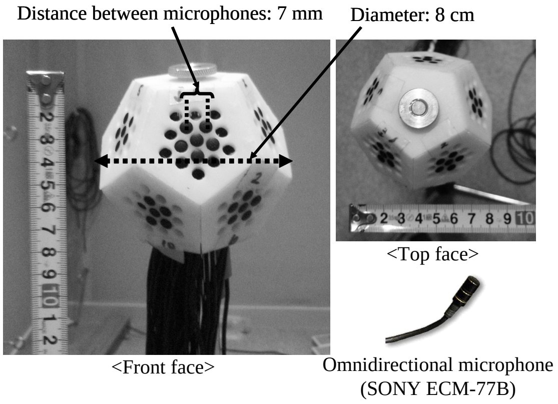

Distance between microphones: 7 mm
Diameter: 8 cm
<Top face>
<Front face>
Omnidirectional microphone
(SONY ECM-77B)

Figure 1: Developed dodecahedral microphone array made from ABS resin. Ten faces except top and bottom are available to install microphones, and maximum number of microphones is 160. Here, six microphones were installed around the center of each face.

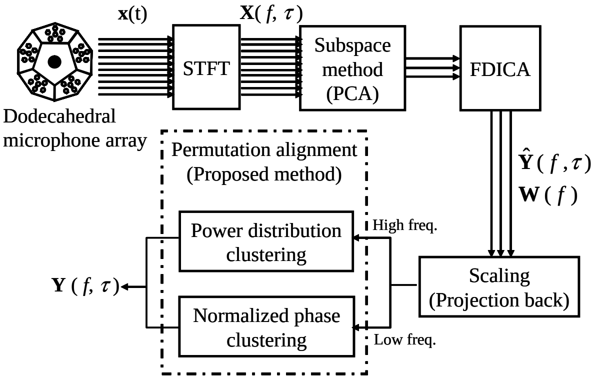

x(t)
X(f, τ)
STFT
Subspace method (PCA)
FDICA
Dodecahedral microphone array
Permutation alignment (Proposed method)
Ŷ(f, τ)
W(f)
Power distribution clustering
High freq.
Y(f, τ)
Scaling (Projection back)
Normalized phase clustering
Low freq.

Figure 2: Block diagram of separation procedure. Our proposed part is shown in the center (permutation alignment scheme).

mm. The role of top and bottom faces is for installing in the microphone stand. Our method solves FD-ICA's permutation problem by using the developed device. The observed signals at each face have different acoustic features such as sound pressure levels, arrival times, influences of diffraction waves, and so on. Therefore, our proposed method uses these features to group the frequency components of the separated signals that are obtained by FD-ICA. Gain and phase information are applied to solve the permutation problem at high and low frequencies, respectively. In addition, since a human being's sound localization queue is considered different between the high and low frequency ranges [8], we also refer to it.

3. SIGNAL SEPARATION USING FREQUENCY-DOMAIN INDEPENDENT COMPONENT ANALYSIS

Figure 2 shows a block diagram of the blind signal separation process with the developed dodecahedral microphone array. Our proposed part is shown in the center, and the other parts employed the method proposed by previous researches [9, 10, 11]. Mixture signal $\mathbf{x}(t)$ is observed by the microphone array. The mixture signals were convolved N source signals with an acoustic transfer function between sound source n and microphone m. A final separated signal $\mathbf{Y}(f,\tau)$ is obtained from an observed signal $\mathbf{X}(f,\tau)$ by FD-ICA. To perform FD-ICA, the dimension of the observed signals is reduced from the number of microphones M to the number of sound source signals N by the subspace method [9]. Separation matrix $\mathbf{W}(f)$ is calculated with the natural gradient algorithm based on Kullback-Leibler (KL) divergence minimize [10], and then separation signals $\hat{\mathbf{Y}}(f,\tau)$ are obtained. Since the FD-ICA method has scaling and permutation problems, we use the projection back method [11] for the scaling problem and the proposed method for the permutation problem. A final separation signal $\mathbf{Y}(f,\tau)$ is obtained by solving the permutation problem.

4. SOLVING PERMUTATION PROBLEM USING DODECAHEDRAL MICROPHONE ARRAY

Gain and phase information are applied to solve the permutation problem at high and low frequencies, respectively.

4.1 Grouping using acoustic pressure distribution at high frequency range

This method uses acoustic pressure distribution $\mathbf{p}$ observed on the surface of the dodecahedral microphone array. These distributions correspond to each source signal. Acoustic pressure distribution $\mathbf{p}$ is obtained by the acoustic pressure at each face $p_{i,l}$. Acoustic pressure $p_{i,l}$ is described by (1):

$$p_{i,l}(f) = \frac{1}{|M(l)|}\sum_{m\in M(l)} |w^{+}_{i,m}(f)|, \quad l = 1,\cdots,10, \qquad (1)$$

where $M(l)$ and $w^{+}(f)$ denote a set of microphones included in the l^{th} face and the transfer function from each source to each microphone calculated by the pseudo-inverse of separation matrix $\mathbf{W}(f)$, respectively. Then, vector $\mathbf{p}$ is calculated by acoustic pressure $p_{i,l}$ at 10 faces where the microphones can be set:.

$$\mathbf{p}_i(f) = [p_{i,1}(f), p_{i,2}(f), \cdots, p_{i,10}(f)], \quad i = 1,\cdots,N, \quad (2)$$

where N is the number of sound sources. Finally, a normalization scheme for vector $\mathbf{p}_i$ is performed:

$$\mathbf{p}_i(f) \leftarrow \frac{\mathbf{p}_i(f)}{\sum_{l=1}^{10} p_{i,l}}. \qquad (3)$$

Grouping is accomplished by k-means algorithm for all frequency acoustic pressure distribution $\mathbf{p}$. The cost function of the grouping is described by (4):

$$\text{Err} = \sum_{k=1}^{N}\sum_{\mathbf{p}\in C_k} ||\mathbf{p} - \mathbf{c}_k||^2, \qquad (4)$$

where C_k represents cluster k whose centroid is $\mathbf{c}_k$. The centroid is calculated with all acoustic pressure distributions (number of frequency bins) × (number of source signals). Then the distances between the centroid and the pressure distribution that correspond to all sources are evaluated for each frequency. Finally, permutation matrix $\Pi(f)$ is estimated:

$$\Pi(f) = \underset{\Pi}{\text{argmin}} \sum_{k=1}^{N} ||\mathbf{p}_{\Pi_k}(f) - \mathbf{c}_k||^2. \qquad (5)$$

These procedures are executed in the high frequency range, where sound-wave damping is large and diffraction is small.

4.2 Grouping using normalized relative phase at low frequency range

Phase information is used for the grouping processes at the low frequency range. In this part, normalized phase feature ϕ is used as the phase information. The normalized phase feature is obtained by the pseudo-inverse of separation matrix $\mathbf{W}(f)$:

$$\phi(\mathbf{w}_q^+(f)) = [\exp(j\tau_{q,1}),\cdots,\exp(j\tau_{q,M})], \quad q = 1,\cdots,N, \tag{6}$$

where $\mathbf{w}^+$ is a row vector of $\mathbf{W}^+$ and + denotes the pseudo-inverse. $\tau_{q,m}$ is the normalized delay given by

$$\tau_{q,m} = \beta\frac{\arg(w_{q,m}^+(f))}{f}, \tag{7}$$

where β is a normalization constant. The permutation problem can be solved by grouping this normalized phase feature. However, the similarity between phase vectors can't be evaluated simply, for example, Euclidean distance, due to phase shift $\exp(j\theta_\varepsilon)$ between two frequency components of the same source, $s_\alpha(f_\psi)$ and $s_\alpha(f_\varphi)$. Therefore, similarity between normalized phase vectors is defined by (8) reference from [2]:

$$\begin{aligned}&\mathrm{Sim}\big(\mathbf{w}_\alpha^+(f_\varphi),\mathbf{w}_\beta^+(f_\psi)\big)\\ &= \sum_{l=1}^{10}\left|\sum_{m\in M(l)}\phi(w_{\alpha,m}^+(f_\varphi))^*\cdot\phi(w_{\beta,m}^+(f_\psi))\right|,\end{aligned} \tag{8}$$

where * denotes the complex conjugate. First, this similarity cost function calculates the conjugate inner product:

$$\phi(w_{\alpha,m}^+(f_\varphi))^*\cdot\phi(w_{\beta,m}^+(f_\psi)), \tag{9}$$

and then the absolute value of the summation inner product is calculated. By calculating the absolute value, this cost function is robust to constant phase shift $\exp(j\theta_\varepsilon)$. Therefore, this grouping method evaluates the relative phase pattern between microphones. In the same way as a high frequency range procedure, the permutation matrix is decided with the k-means algorithm and a cost function. Here, the procedure that updates centroid $\bar{\mathbf{w}}_k^+$ is performed by (10):

$$\begin{aligned}\bar{\mathbf{w}}_k^+ &\leftarrow \frac{1}{Q}\textstyle\sum_{q\ \text{s.t.``}\phi(\mathbf{w}_q^+(f))\in C_k\text{''}}^{Q} I, \\ I &= \{\phi(\mathbf{w}_q^+(f))\cdot\exp(-j\arg((\bar{\mathbf{w}}_k^+)^H\phi(\mathbf{w}_q^+(f))\},\end{aligned} \tag{10}$$

where Q is the number of elements in the k^{th} cluster.

5. EXPERIMENTS

5.1 Experimental conditions

The performances of the proposed method were evaluated by sound source separation experiments. Test signals were generated by the convolution of sound source signals and impulse responses between a loudspeaker (BOSE ACOUSTIMASS) and omni-directional microphones (SONY ECM-77B). Speech and instrumental signals were used for sound source signals. We evaluated two conditions of speech signals (male and female speech) and a condition of instruments (drums, guitar, and bass) shown by Table 1. Speech sets 1 and 2 consisted of different phrases. The number of sound sources were given, and the locations were unknown. Experiments were performed in a soundproof chamber whose reverberation time was 138 msec. The other experimental conditions are shown in Table 2.

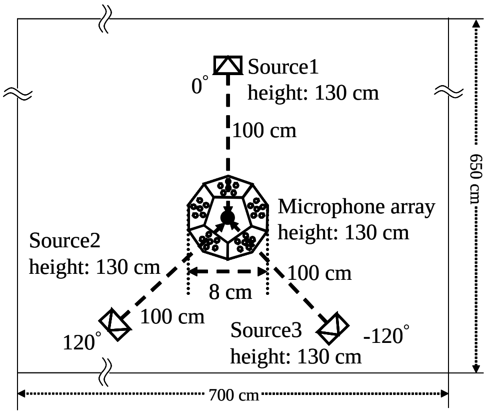


Figure 3: Experimental setup. Room reverberation time is 138 msec.

Table 1: Test set

Speech set 1	Source 1: Female Source 2: Female Source 3: Male
Speech set 2	Source 1: Female Source 2: Female Source 3: Male
Instruments set	Source 1: Drums Source 2: Bass Source 3: Guitar

5.2 Results

In our experiments, the high frequency range was from 4 to 8 kHz, and the low frequency range was from 0 Hz to 4 kHz. Grouping processes were respectively performed in the high and low frequency ranges, and the resultant output signals were combined by hand. Separation performances were evaluated by an improvement of the signal-to-interference ratio (SIR) given by (11):

$$\text{SIR improvement}_n = \text{OutputSIR}_n - \text{InputSIR}_n \quad [\text{dB}], \tag{11}$$

$$\text{InputSIR}_n = 10\log_{10}\left[\frac{\sum_t x_{mn}(t)^2}{\sum_t\{\sum_{s\neq n}x_{ms}(t)\}^2}\right] \quad [\text{dB}], \tag{12}$$

$$\text{OutputSIR}_n = 10\log_{10}\left[\frac{\sum_t y_{nn}(t)^2}{\sum_t\{\sum_{s\neq n}y_{ns}(t)\}^2}\right] \quad [\text{dB}], \tag{13}$$

where x_{ms} is an input signal from source signal s observed by microphone m and y_{ns} is an output signal from source signal n processed by separation filter w_s.

Figure 4 shows the grouping result in the high frequency range when speech signals were used. Dotted lines denote cluster centroids. Similar acoustic pressure distributions

Table 2: Experimental conditions

Sampling frequency	16 kHz
Length of frame	1024 pt (64 msec)
Frame shift	256 pt (16 msec)
Window function	Hamming
Length of STFT	1024 pt
Background noise level	10.7 dB(A)
Sound pressure level (1m)	88.3 dB(A)
Temperature	13.7 °C
Number of microphones	60
Number of sources	3

could be assembled by the grouping method. Figure 5 shows one result of the inner product in the low frequency range. The conjugate inner product is plotted between one normalized phase feature at 781 Hz (the 50^{th} bin) and three centroids. The absolute values of the summation of this inner product are compared, and this frequency component was clustered to the third cluster. Similarities among all frequency features and three centroids are shown in Figure 6. In each cluster, high similarity components existed in each frequency and were grouped.

The results were compared with the ideal condition and the conventional method [7]. In the case of the ideal condition, the permutation problem was solved since the sound source signals were known. Therefore, the highest performance is obtained by the ideal condition. The conventional method uses time delays and differences of sound attenuation that were observed among a sound source and microphones. This conventional method groups phase and amplitude normalized vector $\bar{\mathbf{a}}_r(f) = [\bar{a}_{r,m}(f), \ldots, \bar{a}_{r,M}(f)]$:

$$\bar{a}_{r,m}(f) = |w^+_{r,m}(f)| \exp\left[j \frac{\arg[w^+_{r,m}(f)/w^+_{r,J}(f)]}{4fc^{-1}d_{max}} \right], \quad (14)$$

where J and d_{max} are the index numbers of the reference microphone and the constant value, for example, the maximum distance between microphones. Then normalized vector $\bar{\mathbf{a}}(f)$ is grouped by the k-means algorithm. The cost function is described by (15):

$$\mathrm{Err} = \sum_{k=1}^{N} \sum_{\bar{\mathbf{a}} \in C_k} ||\bar{\mathbf{a}} - \mathbf{c}_k||^2. \quad (15)$$

Table 3 shows the SIR improvement score. Figures 7 and 8 show the spectrogram of female speech and bass signals, respectively. In both figures, the spectrograms of the mixture, separated, and source signals are shown. The separation performances obtained by the proposed method outperformed the conventional method. The separation performances of the speech sets were especially close to the ideal condition, and the average SIR improvement was more than 20 dB. The proposed method has an advantage over the conventional method due to dividing the frequency range. The phase and amplitude information mutually interfered with the conventional method. However, the performance of the instruments was poor. In Figure 8, interference signals were caused by the separation error. This failure occurred by the differences of the dominant frequency range among the instruments. For example, the dominant frequency range of the bass is from dozens to hundreds of Hz and the drums is wide frequency range. Therefore, reflections or noises caused a mistake of the supposed number of sources estimated by FD-ICA. Applying the subspace method must be improved.

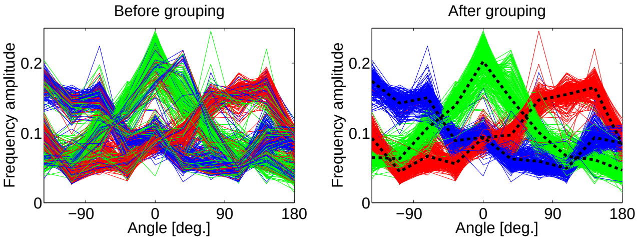


Figure 4: Grouping result of high frequency (4-8 kHz) using acoustic pressure distribution. Dotted lines denote cluster centroids.

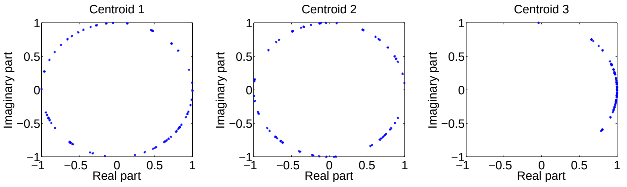


Figure 5: Example of inner product in low frequency (0 Hz-4 kHz). Conjugate inner products between normalized phase feature at 50^{th} bin (781 Hz) and three centroids are plotted. Absolute values of the summation of this inner product are compared, and this frequency component is clustered to the third cluster.

6. SUMMARY AND FUTURE WORKS

In this paper, a small dodecahedral microphone array was developed and a grouping method of frequency components for FD-ICA using the developed device was proposed. The proposed method uses an acoustic pressure distribution that observed the faces of the device and normalized the relative phases at each microphone in high and low frequency ranges, respectively. The experimental results showed that the SIR improvement score of the proposed method was more than 20 dB in the case of speech signals. Moreover, the proposed method was better than the conventional method and close to the ideal conditions. However, the performances were poor in the case of the instruments. Future work includes improving the estimation method of the number of sound sources and developing a method of synthesizing separated signals at high and low frequency ranges.

REFERENCES

[1] T. Fujii and M. Tanimoto, “Free-viewpoint TV system based on the ray-space representation,” *SPIE ITCom*, vol. 4864-22, pp. 175-189, 2002.

[2] K. Niwa, T. Nishino, and K. Takeda, “Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation,” *ICASSP2008*, pp. 181-184, 2008.

[3] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, “A realtime multimodal

Table 3: SIR improvement score [dB]

	Speech set1			Speech set2			Instruments set		
	Female	Female	Male	Female	Female	Male	Drums	Bass	Guitar
Proposal method	27.1	28.2	21.5	22.0	33.8	25.2	18.6	3.5	15.6
Conventional method	25.1	22.5	17.4	19.0	30.7	21.1	17.2	3.4	14.6
Ideal condition	27.4	30.1	23.5	23.0	34.5	26.3	22.5	6.0	27.5

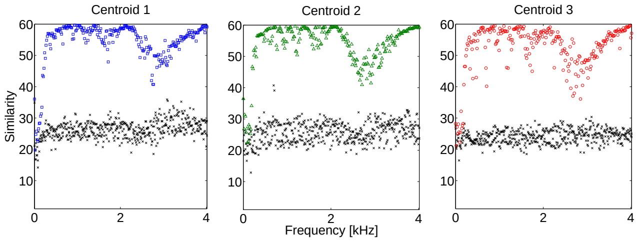


Figure 6: Example of clustering in low frequency (0 Hz - 4 kHz). Similarities between all frequency features and three centroids are shown. In each cluster, high similarity components (color symbols) existed in each frequency and were grouped.

system for analyzing group meetings by combining face pose tracking and speaker diarization," *ICMI'08*, pp. 257-264, 2008.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21-34, 1998.

[5] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," *ICANN'98*, pp. 761-766, 1998.

[6] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *ICASSP2002*, pp. 881-884, 2002.

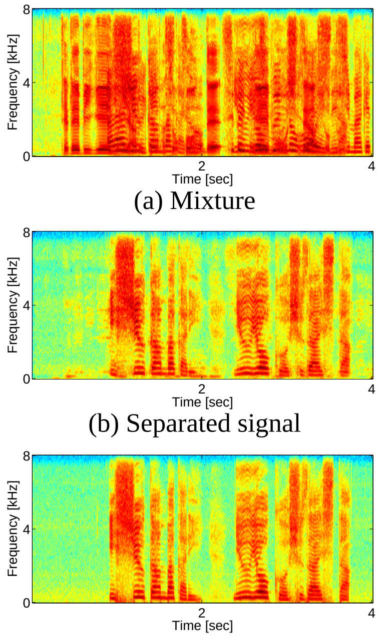
[7] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and Time-Frequency Masking," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165-2173, 2006.

[8] J. Blauert, *Spatial hearing (revised edition)*, The MIT Press, 1996.

[9] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387-392, 1985.

[10] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, New York Wiley, 2001.

[11] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *International Symposium on Nonlinear Theory and Its Application*, vol. 3, pp. 923-926, 1998.

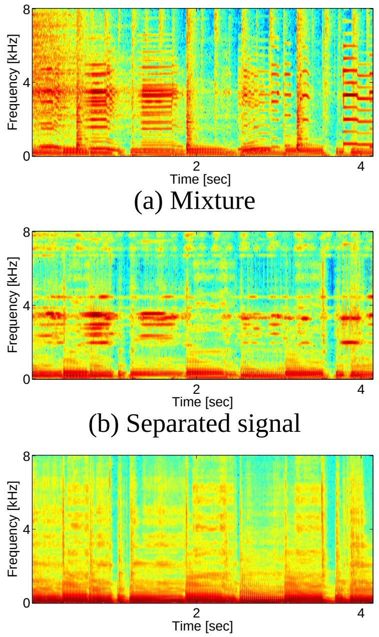


(a) Mixture

(b) Separated signal

(c) Source signal

Figure 7: Spectrogram of mixture, separated, and source signals of female speech (source 1 of speech set 1).



(a) Mixture

(b) Separated signal

(c) Source signal

Figure 8: Spectrogram of mixture, separated, and source signals of bass signal (instrument set).