BLIND SOURCE SEPARATION ON IPHONE IN REAL ENVIRONMENT

Nobutaka Ono

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
onono@nii.ac.jp

ABSTRACT

A stereo blind source separation system for the iPhone is presented. A commercially available stereo microphone is installed on an iPhone and an input sound can be recorded by it or be transferred by PC through iTunes. The sound is then separated into two sources by a stereo-specific version of auxiliary-function-based independent vector analysis (AuxIVA2) [1] in 1/5 the computational time of the input signal length. The quantitative evaluation for speech separation and other examples for source separation in real environments are presented.

Index Terms— Blind source separation, independent vector analysis, auxiliary function, mobile phone

1. INTRODUCTION

Blind source separation (BSS) is a technique to extract a desired source from mixtures that is often used for speech recognition, communication, acoustic event detection, and scene analysis. For convolutive overdetermined mixtures, independent component analysis (ICA) followed by permutation correction [2] or independent vector analysis (IVA) [3, 4, 5] in the frequency domain have been developed this decade as standard techniques. However, reducing the computation time remains an issue [7], especially for mobile phone or hearing aid applications due to the limitations of CPU performance and batteries.

The author recently developed a fast and stable algorithm for ICA [8] and IVA [9][10] based on the auxiliary function technique. Its faster version specifically for stereo case and its implementation on an iPhone were also presented [1]. In this paper, an overview of auxiliary-function-based independent vector analysis for stereo case (AuxIVA2), its implementation as an iPhone application, and an evaluation of computational time and separation performance are described.

2. FREQUENCY-DOMAIN BLIND SOURCE SEPARATION

Assume here that K sources are observed by K microphones and that their short-time Fourier transform (STFT) representations are obtained. Let $\boldsymbol{s}(\omega,\tau)$, $\boldsymbol{x}(\omega,\tau)$, and $\boldsymbol{y}(\omega,\tau)$ be the frequency-wise vector representation of the sources, the observations, and the estimated sources, respectively, which are defined as

$$\boldsymbol{s}(\omega,\tau) = (s_1(\omega,\tau)\ \cdots\ s_K(\omega,\tau))^t, \tag{1}$$

$$\boldsymbol{x}(\omega,\tau) = (x_1(\omega,\tau)\ \cdots\ x_K(\omega,\tau))^t, \tag{2}$$

$$\boldsymbol{y}(\omega,\tau) = (y_1(\omega,\tau)\ \cdots\ y_K(\omega,\tau))^t, \tag{3}$$

where ^{t} denotes the vector transpose and the size of each vector is $K \times 1$. In the frequency-domain approach for a convolutive mixture, a linear mixing model,

$$\boldsymbol{x}(\omega,\tau) = A(\omega)\boldsymbol{s}(\omega,\tau), \tag{4}$$

is assumed, where $A(\omega)$ is a $K \times K$ mixing matrix. The sources are estimated by a linear demixing process,

$$\boldsymbol{y}(\omega,\tau) = W(\omega)\boldsymbol{x}(\omega,\tau), \tag{5}$$

where $W(\omega) = (\boldsymbol{w}_1(\omega)\ \cdots\ \boldsymbol{w}_K(\omega))^h$ is a $K \times K$ demixing matrix and ^{h} denotes conjugate transpose.

3. OVERVIEW OF AUXIVA

3.1. Objective Function of IVA

In IVA, assuming a multivariate p.d.f. for sources to exploit the dependencies over frequency components, the demixing matrices are estimated by minimizing the following objective function.

$$J(\boldsymbol{W}) = \sum_{k=1}^{K} \frac{1}{N_\tau} \sum_{\tau=1}^{N_\tau} G(\boldsymbol{y}_k(\tau)) - \sum_{\omega=1}^{N_\omega} \log|\det W(\omega)|, \tag{6}$$

where $\boldsymbol{W}$ denotes a set of $W(\omega)$, N_ω and N_τ are the number of frequency bins and time frames, respectively, $\boldsymbol{y}_k(\tau)$ is the source-wise vector representation defined as

$$\boldsymbol{y}_k(\tau) = (y_k(1,\tau)\ \cdots\ y_k(N_\omega,\tau))^t, \tag{7}$$

This research was partially supported by Grand-in-Aid for Scientific Research (A) (23240023) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The author acknowledges Redec Co., Ltd. for contribution to development of the presented AuxIVA2 on iPhone.

and $G(\boldsymbol{y}_k(\tau))$ is called a contrast function. In the literature [3, 4, 5], spherical contrast functions,

$$G(\boldsymbol{y}_k(\tau)) = G_R(r_k(\tau)) \tag{8}$$

$$r_k(\tau) = ||\boldsymbol{y}_k(\tau)||_2 = \sqrt{\sum_{\omega=1}^{N_\omega} |y_k(\omega,\tau)|^2}, \tag{9}$$

are often used, where $||\cdot||_2$ denotes the L_2 norm of a vector. In AuxIVA, $G_R(r)$ has to be selected such that $G'_R(r)/r$ is monotonically decreasing in $r > 0$ [9]. A typical choice is $G_R(r) = r$, which corresponds to the Laplace distribution.

3.2. Auxiliary function approach to IVA

Minimizing eq. (6) is a nonlinear optimization problem and there are generally no closed-form solutions. The standard approach to solve it is iteratively applying the gradient-based update rule [3, 4, 5]. However, there is a tradeoff between the convergence speed and the stability, which are dependent on the step-size parameter.

In contrast, in AuxIVA, instead of directly decreasing eq. (6), the demixing matrix is estimated by alternatively calculating and minimizing the following auxiliary function.

$$Q(\boldsymbol{W}, \boldsymbol{r}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{\omega=1}^{N_\omega} \boldsymbol{w}_k^h(\omega) V_k(\omega) \boldsymbol{w}_k(\omega) - \sum_{\omega=1}^{N_\omega} \log|\det W(\omega)| + R, \tag{10}$$

where $\boldsymbol{r}$ denotes a set of auxiliary variables, $r_k(\tau)$, which are included in $V_k(\omega)$, and R represents a constant term independent of $\boldsymbol{W}$. As with calculating and minimizing Q-function in the EM algorithm, the monotone decrease of the objective function is guaranteed.

4. FAST ALGORITHM FOR STEREO INDEPENDENT VECTOR ANALYSIS

In AuxIVA, auxiliary variable updates and demixing matrix updates are alternatively applied and the demixing matrix is updated by minimizing the auxiliary function written in eq. (10). In general cases (e.g., $K \geq 3$), closed-form solutions have never been found for $\partial Q(\boldsymbol{W}, \boldsymbol{r})/\partial W(\omega) = 0$. Hence, in AuxIVA, each row vector of the demixing matrix, $\boldsymbol{w}_k$, is updated in order by minimizing $Q(\boldsymbol{W}, \boldsymbol{r})$ in terms of $\boldsymbol{w}_k$ with other $\boldsymbol{w}_l (l \neq k)$ fixed [9]. However, in stereo case (e.g., $K = 2$), $\partial Q(\boldsymbol{W}, \boldsymbol{r})/\partial W(\omega) = 0$ becomes equivalent to a generalized eigenvalue problem and can therefore be solved in a closed-form. This is referred to as AuxIVA2, which has a faster convergence than general AuxIVA [1]. The algorithm of AuxIVA2 is summarized as follows. Auxiliary variable updates and demixing matrix updates are alternatively applied for all ω.

Auxiliary variable updates
Eq. (11) is first calculated for both $k = 1$ and $k = 2$. Eq. (12) is then calculated for all frequency bins.

$$r_k(\tau) = \sqrt{\sum_{\omega=1}^{N_\omega} |\boldsymbol{w}_k^h(\omega)\boldsymbol{x}(\omega,\tau)|^2} \tag{11}$$

$$V_k(\omega) = \frac{1}{N_t}\sum_{t=1}^{N_t}\left[\frac{G'_R(r_k(t))}{r_k(t)}\boldsymbol{x}(\omega,\tau)\boldsymbol{x}^h(\omega,\tau)\right] \tag{12}$$

Demixing matrix updates

$$H(\omega) = V_1^{-1}(\omega)V_2(\omega), \tag{13}$$

$$\lambda_1(\omega) = \frac{\mathrm{tr}(H(\omega)) + \sqrt{\mathrm{tr}(H(\omega))^2 - 4\mathrm{det}(H(\omega))}}{2}, \tag{14}$$

$$\lambda_2(\omega) = \frac{\mathrm{tr}(H(\omega)) - \sqrt{\mathrm{tr}(H(\omega))^2 - 4\mathrm{det}(H(\omega))}}{2}, \tag{15}$$

$$\boldsymbol{e}_1(\omega) = \begin{pmatrix} H_{22}(\omega) - \lambda_1(\omega) \\ -H_{21}(\omega) \end{pmatrix}, \tag{16}$$

$$\boldsymbol{e}_2(\omega) = \begin{pmatrix} -H_{12}(\omega) \\ H_{11}(\omega) - \lambda_2(\omega) \end{pmatrix}, \tag{17}$$

$$\boldsymbol{w}_1(\omega) = \frac{\boldsymbol{e}_1(\omega)}{\sqrt{\boldsymbol{e}_1^h(\omega)V_1(\omega)\boldsymbol{e}_1(\omega)}}, \tag{18}$$

$$\boldsymbol{w}_2(\omega) = \frac{\boldsymbol{e}_2(\omega)}{\sqrt{\boldsymbol{e}_2^h(\omega)V_2(\omega)\boldsymbol{e}_2(\omega)}}, \tag{19}$$

where $H_{ij}(\omega)$ denotes the ijth element of $H(\omega)$.

5. IMPLEMENTATION ON IPHONE

5.1. StereoSep: iPhone application for stereo blind source separation

As a prototype stereo BSS system on a mobile phone, we implemented the AuxIVA2 as an iPhone application, called "StereoSep", in cooperation with Redec Co., Ltd. In this implementation, a vector library called vDSP on Mac OS X [12] was exploited as much as possible for fast computation. $G_R(r) = r$ was used as the contrast function.

A photograph of an iPhone with commercially available stereo microphones is shown in Fig. 1 and the home screen and setting screen of the application are shown in Fig. 2. This application is only applicable for a stereo signal, which can be recorded by the "REC" button [1] or can be selected by the "SELECT" button from a list of WAV files transferred from a PC through iTunes in advance. In addition, the sampling frequency for recording, the frame length, and the iteration number of AuxIVA2 can be changed in the setting screen.

[1] Because the iPhone itself has a single microphone for recording, this function is possible only if a stereo microphone is installed on it.

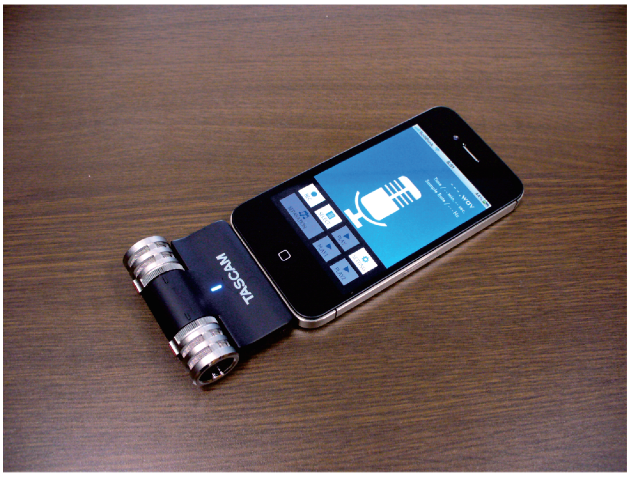
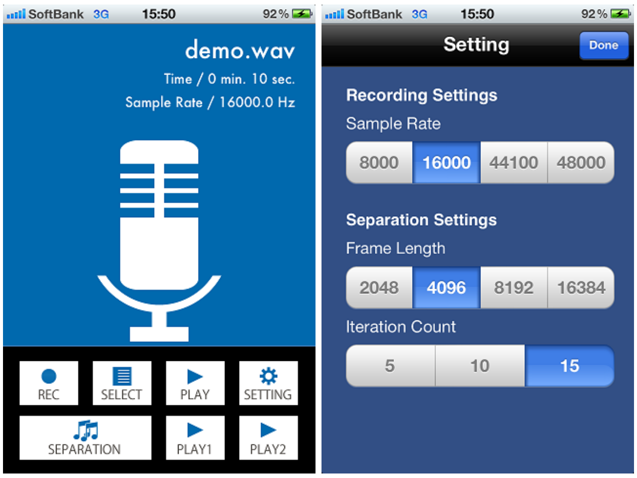


Fig. 1. Photo of an iPhone 4 with stereo microphone (TASCAM iM2 provided by TEAC corporation).



demo.wav

Time / 0 min. 10 sec.

Sample Rate / 16000.0 Hz

Setting

Recording Settings

Sample Rate: 8000 [✓ 16000] 44100 48000

Separation Settings

Frame Length: 2048 [✓ 4096] 8192 16384

Iteration Count: 5 10 [✓ 15]

Fig. 2. The main (left) and setting (right) screen images of the stereo BSS iPhone application.

5.2. Evaluation of computation time

Figure 3 shows the evaluation results of the computation time for separation on an iPhone 4 to a different-length input signal, where a 4096-point frame length was used.

As shown in the figure, the computation time is almost linear to the length of the input signal. While, it is not linear to the number of iterations because the computation includes not only iterative estimates of demixing matrices but also STFT, the projection back for adjusting the scale, and the inverse STFT. Because the computation times for 5, 10, and 15 iterations are about $5 : 7 : 9$, the computation time of such overhead is estimated to be almost that of 7.5 iterations $(5 + 7.5 : 10 + 7.5 : 15 + 7.5 = 5 : 7 : 9)$.

The computation time does not depend on the frame length so much. For an 8192-point frame length, it is just 3 to 10% larger than the 4096-point case.

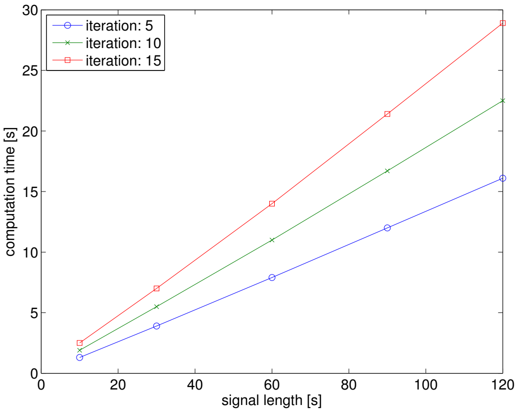


Fig. 3. The relationship between input signal length and calculation time on an iPhone 4.

6. EXPERIMENTAL EVALUATION OF SEPARATION

6.1. Quantitative evaluation of separation performance

The setup of four loudspeakers and an iPhone is shown in Fig. 4. The reverberation time was about 400 ms. The developed iPhone app could separate the directly recorded mixture of two sources. However, for the quantitative evaluation of the separation performance, we need the ground-truth source image. We therefore conducted the experiments as follows.

From each loudspeaker (A, B, C, and D), 10s-length different speech selected from development data (dev1) in the underdetermined-speech and music mixture task in SiSEC 2008 [13] were played. First, each source image was recorded on an iPhone with a stereo microphone and was transferred to a PC. Then, mixtures of two of them in all combinations were obtained by summing them on the PC. They were then transferred back to the iPhone. Finally, source separation was applied on the iPhone. The iteration number was set to 15 and the sampling frequency was 16 kHz. The separation performance was evaluated on the PC with Signal to Distortion Ratio (SDR) and Signal to Interference Ratio (SIR) criterions provided by the BSSeval Toolbox [11].

The separation results are shown in Fig. 5. Because the magnitudes of sources were not controlled, the combinations of sources included some magnitude mismatch. For example, in the mixture of source B and source C, the original SIR (before separation) was ± 12 dB because of the smaller magnitude of source C. However, in most cases, source separation worked well. We can see a dependency of the appropriate frame length on the distance from sources. For closer sources (mixing A and B), a shorter frame length showed a better performance, while for further sources (mixing C and D), the longer the frame length the better. In average for all combinations, 7.5-dB improvement in SIR was achieved with the 4096-point frame length.

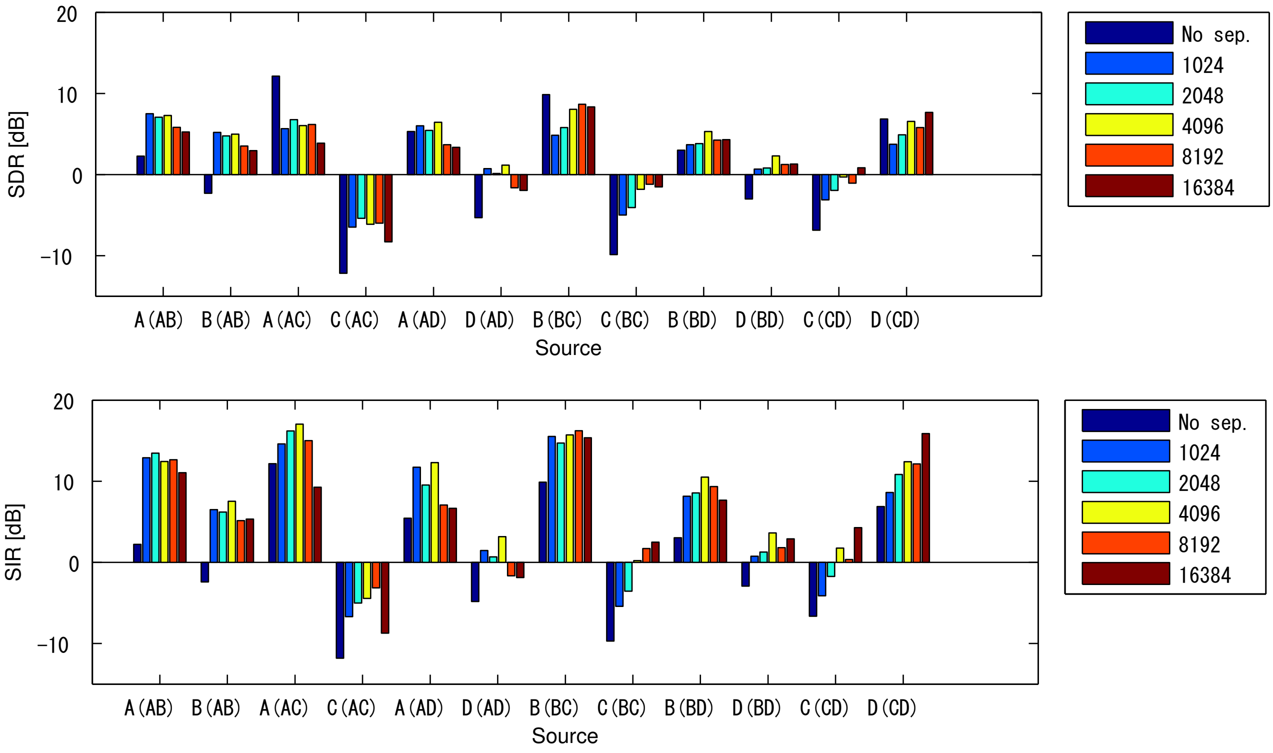


Fig. 5. Separation performance with SDR (top) and SIR (bottom) criterions for two-speech separation. "No sep." denotes before separation while "1024", "2048", etc. denote the frame length used for separation. In the horizontal axis, A(AB) denotes source A in the mixture of source A and source B.

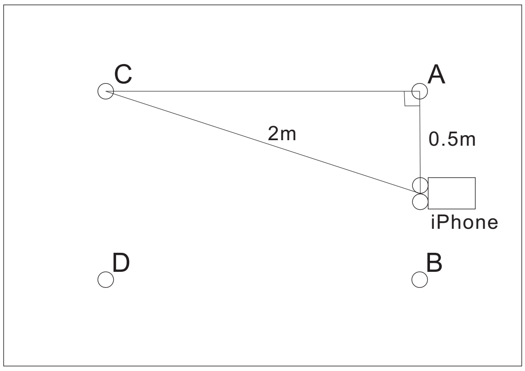


Fig. 4. Experimental setup of loudspeakers and microphones.

6.2. Other separation examples

The implementation of source separation on an iPhone facilitates investigating the applicability of source separation to various sounds in the real world. One typical example is shown in Fig. 6, where female speech was present with the sound of a news program from TV. The female speaker and the TV were located about 1.5 m from the iPhone and were 120 degrees apart. The sampling frequency was 16 kHz. Only 0-2 kHz and a 6-s fragment is displayed in the figure. The two sounds were well separated under the 4096-point frame length condition. Another example is shown in Fig. 7, where two announcements on different platforms in a subway station announcing the approach of a train were overlapped. Because the subway station was reverberant and the distance between the loudspeakers and the iPhone was several meters, separation with the 16,384-point frame length was not perfect, but its effect can be heard clearly.

7. CONCLUSION

In this paper, a stereo blind source separation system for the iPhone is presented. The stereo-specific version of auxiliary-function-based independent vector analysis (AuxIVA2) enables fast and stable separation. The quantitative evaluation for speech separation and other examples for source separation in real environments are presented.

8. REFERENCES

[1] N. Ono, "Fast Stereo Independent Vector Analysis and its Implementation on Mobile Phone," *Proc. IWAENC*, Sept. 2012.

[2] H. Sawada, R. Mukai, S. Araki, S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation,"*IEEE Trans. SAP*, vol. 12, no. 5, pp. 530-538, 2004.

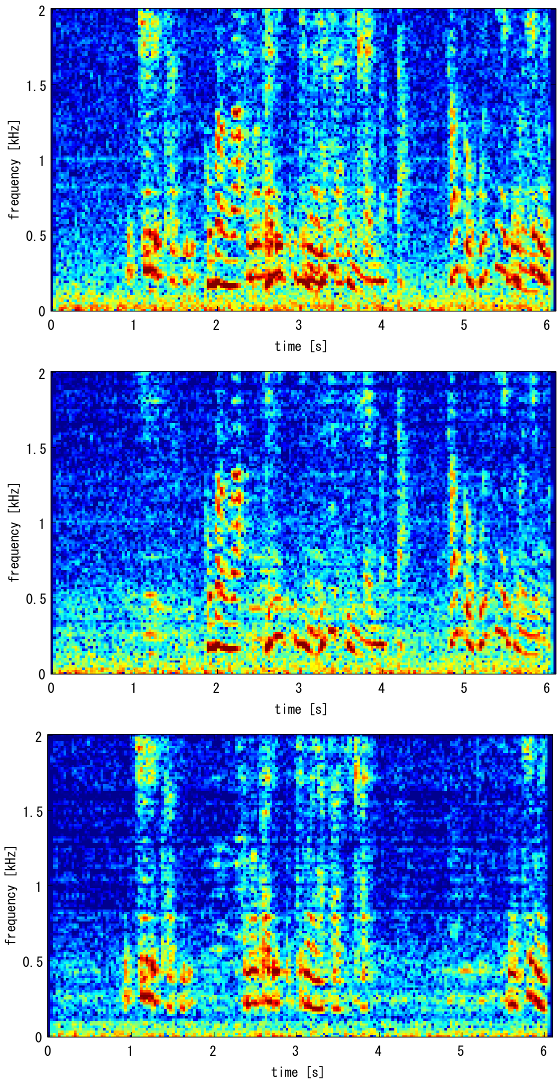


Fig. 6. Spectrograms of news program from TV + speech (top), the separated news program (middle), and the separated speech (bottom).

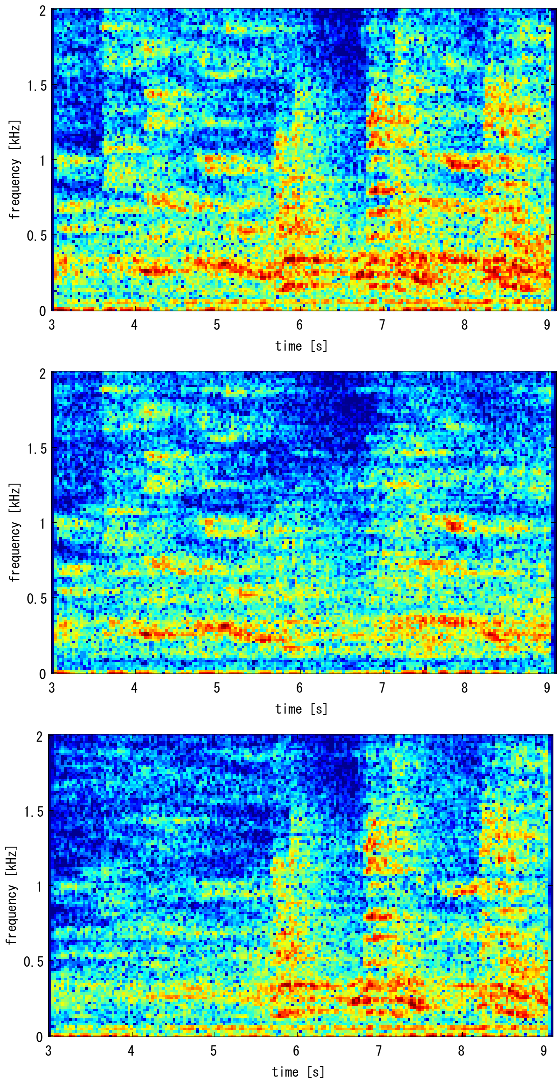


Fig. 7. Spectrograms of two overlapped announcements in a metro station (top), the separated further announcement (middle), and the separated closer announcement (bottom).

[3] A. Hiroe, “Solution of Permutation Problem in Frequency Domain ICA Using Multivariate Probability Density Functions,” *Proc. ICA*, pp. 601–608, 2006.

[4] T. Kim, T. Eltoft, and T.-W. Lee, “Independent Vector Analysis: An Extension of ICA to Multivariate Components,” *Proc. ICA*, pp. 165–172, 2006.

[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind Source Separation Exploiting Higher-order Frequency Dependencies,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[6] P. Smaragdis, “Blind Separation of Convolved Mixtures in the Frequency Domain,” *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[7] K. Osako, Y. Mori, Y. Takahashi, H. Saruwatari, and K. Shikano, “Fast Convergence Blind Source Separation based on Frequency Subband Interpolation by Null Beamforming,” *Proc. WASPAA*, pp. 42-45, Oct. 2007.

[8] N. Ono and S. Miyabe, “Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources,” *Proc. LVA/ICA*, pp.165-172, 2010.

[9] N. Ono, “Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique,” *Proc. WASPAA*, pp. 189-192, Oct. 2011.

[10] N. Ono, “Auxiliary-function-based Independent Vector Analysis with Power of Vector-norm Type Weighting Functions,” *Proc. APSIPA*, Dec. 2012.

[11] E. Vincent, C. Fevotte, and R. Gribonval, “Performance Measurement in Blind Audio Source Separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[12] https://developer.apple.com/library/mac/#documentation/ Accelerate/Reference/vDSPRef/Reference/reference.html

[13] http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Underdetermined+speech+and+music+mixtures