INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Multi-talker Speech Recognition Based on Blind Source Separation with Ad hoc Microphone Array Using Smartphones and Cloud Storage

Keiko Ochi[1], *Nobutaka Ono*[12], *Shigeki Miyabe*[3], *Shoji Makino*[3]

[1]National Institute of Informatics, Tokyo, Japan
[2]SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan
[3]University of Tsukuba, Tsukuba, Japan
{ochi,onono}@nii.ac.jp, {miyabe,maki}@tara.tsukuba.ac.jp

Abstract

In this paper, we present a multi-talker speech recognition system based on blind source separation with an ad hoc microphone array, which consists of smartphones and cloud storage. In this system, a mixture of voices from multiple speakers is recorded by each speaker's smartphone, which is automatically transferred to online cloud storage. Our prototype system is realized using iPhone and Dropbox. Although the signals recorded by different iPhones are not synchronized, the blind synchronization technique compensates both the differences in the time offset and the sampling frequency mismatch. Then, auxiliary-function-based independent vector analysis separates the synchronized mixture into each speaker's voice. Finally, automatic speech recognition is applied to transcribe the speech. By experimental evaluation of the multi-talker speech recognition system using Julius, we confirm that it effectively reduces the speech overlap and improves the speech recognition performance.

Index Terms: Ad hoc microphone array, blind source separation, synchronization, speech recognition

1. Introduction

Multi-party conversation such as at a meeting or a daily chatting is an important target in the automatic speech recognition (ASR) field [1]–[3] because it is a fundamental activty of human communnication. In ASR in a multi-talker environment, the overlap of speech sounds causes problems [4]. The detection of speech overlaps was studied to avoid degrading the performance of ASR [5]. In the Pascal Speech Separation Challenge [6], recognizing a target speech in the presence of another talker's speech was evaluated in a monaural scenario. A multichannel approach has also been studied [7]–[11] because it is more effective for speech separation.

Meanwhile, smartphones are becoming widely prevalent and many people have their own smartphone. Thus, using smartphones as elements of a microphone array may be considered as a promising approach to speech separation. This approach has been studied in the framework of an ad hoc microphone array [12]–[14], which has been applied to beamforming [15]–[17] and blind source separation (BSS) [18]–[20]. It is also attractive for multi-talker speech recognition because it is easy to record conversations without preparing a specific dedicated recording device. The flexibility of the device arrangement makes it easier to set each device close to each talker, which could contribute to improving the input signal-to-noise ratio (SNR).

In an ad hoc microphone array, the signals recorded by different devices are not synchronized because of the different start times of recording and the sampling frequency mismatch between individual devices. In this case, conventional microphone array processing does not work without modification. To overcome this problem, we have developed a blind synchronization technique [21]–[23], which compensates both the difference in the time offset and the sampling frequency mismatch. In our previous work, we showed by simulative experiments that this technique is effective for improving the source separation performance. However it is still challenging to apply this technique to the ASR scenario.

In this paper, we present a multi-talker speech recognition system using smartphones and cloud storage. In this system, a conversation by multiple speakers is recorded by their own iPhones. After recording, an iPhone app that we developed auomatically transfers each recorded signal to fixed Dropbox storage. The collected signals are synchronized by blind synchronization [21]–[23], then auxiliary-function-based independent vector analysis (AuxIVA) [24] is applied for BSS. Finally, each separated signal is transcribed by ASR. By performing experiments using three iPhones with a mixture of three speakers' voices, we confirm that this system effectively reduces the speech overlap and improves the speech recognition performance.

2. Ad hoc microphone array using smartphone and cloud storage

2.1. System Overview

Fig. 1 illustrates an overview of the proposed system. Suppose a situation that multiple speakers have a meeting and they each record their conversation by their smartphones. It is assumed that each device is placed closer to each speaker. The recorded data are individually transferred to online cloud storage by 3G or a Wi-Fi wireless channel. Then, they are synchronized and separated into each speaker's voice in a blind manner, and finally ASR is applied for transcription. All of the signal processing is performed on a PC in a centralized way.

2.2. Prototype Realization Using iPhones and Dropbox

As a realization of the ad hoc microphone array mentioned in the previous subsection, we demonstrate a prototype system using iPhones and Dropbox.

To capture sound, we have developed an iPhone app as shown in Fig. 2. Sound is recorded as an uncompressed PCM-format WAV file with 16 bit precision. To distinguish the files recorded by different devices, the WAV file name has an iden-

http://dx.doi.org/10.21437/Interspeech.2016-758

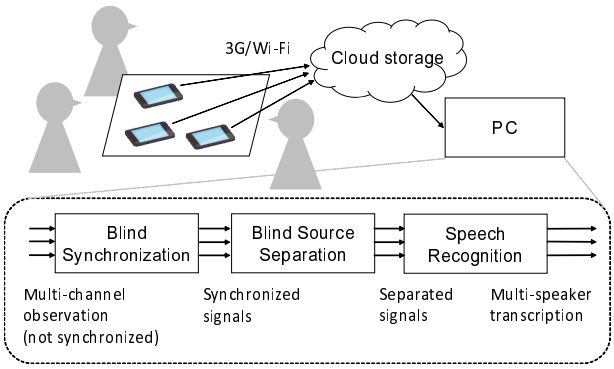

Figure 1: *System overview.*

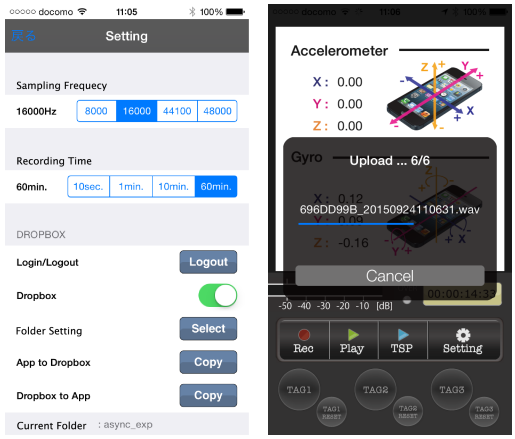

Figure 2: Screenshot of the setup screen (left) and the recording screen during uploading of data to Dropbox (right)

tical device-dependent prefix. Automatic gain control is turned off during recording. One important issue in ad hoc microphone arrays is how to collect or share the data recorded by each device. For example, it would be cumbersome if we have to connect each iPhone with a PC one by one for data transfer. To simplify data transfer, our iPhone app automatically uploads the recorded sound file to fixed Dropbox storage after the recording. This iPhone app records not only sound but also other sensor data such as accelerations, angular accelerations, orientation, and GPS information, which are not used in this work.

3. Signal processing for asynchronous recording of mixture

3.1. Blind Synchronization

For simplicity, we here consider speech captured by two iPhones. Let $x_i(t)$ for $i = 1, 2$ denote the acoustic signals in the continuous time domain at the microphones of the two iPhones. They are independently discretized by the A/D converter (ADC) of each iPhone. Then, because of the different time offsets and the sampling frequency mismatch, the discretized signals are not synchronized. Let $x_i[n]$ for $i = 1, 2$ be the discretized signals, which are given by

$$x_1[n] = x_1\left(\frac{n}{f_s}\right), \tag{1}$$

$$x_2[n] = x_2\left(\frac{n}{(1+\epsilon) f_s} + T_{21}\right), \tag{2}$$

where f_s is the nominal sampling frequency, ϵ is a dimensionless scalar representing the sampling frequency mismatch, the origin of the continuous time $t = 0$ is defined as the time when the sampling of $x_1[n]$ starts, and T_{21} is the continuous time when the sampling of $x_2[n]$ starts.

For conventional array signal processing, synchronization between channels is necessary. Not only the time offset T_{21}, but also the sampling frequency mismatch ϵ greatly degrades the performance of array signal processing [23][25]–[27] even though it can be as small as 10^{-5} depending on the length of the signal. Thus, we have to estimate the synchronized version of the second channel relative to the first channel, which is given as

$$\begin{aligned}\hat{x}_2[n] &= x_2\left(\frac{n}{f_s}\right) \\ &= x_2\left[(1+\epsilon)(n - D_{21})\right],\end{aligned} \tag{3}$$

where $D_{21} = f_s T_{21}$.

To obtain this, a blind synchronization technique [21]–[23] that we have previously developed is applied. In this method, the time offset is first compensated using the cross-correlation of the two observed signals. Then, based on the approximation that the drift of the time difference within a time frame is constant, the sampling frequency mismatch is compensated by a linear phase shift which is given by

$$\hat{X}_2(k, m; \epsilon) = X_2(k, m) \exp\left(\frac{2\pi \jmath k \epsilon m}{L}\right) \tag{4}$$

where $X_2(k, m)$ and $\hat{X}_2(k, m; \epsilon)$ are the STFT representations of $x_2[n]$ before and after the compensation, respectively, k is the discrete frequency index and m denotes the center position of a window with length L in the discrete time (not the discrete frame index). The sampling frequency mismatch ϵ is estimated by maximizing the likelihood of the stationary spatial model in the STFT domain based on the assumption that the sources are motionless and have stationary amplitudes. The estimation of the time offset, the estimation of the sampling frequency mismatch, and reframing using them can be iteratively performed to improve the accuracy. The procedure is described in detail in [23].

3.2. Blind Source Separation

In the supposed situation, the positions of the smartphones are not known in advance. In addition, the blind synchronization technique may change the relationship between the time difference of arrival (TDOA) and the speaker's position. Therefore, BSS is suitable for separating them. In this study, we utilize AuxIVA [24] based on a time-varying Gaussian source model [28] with a small modification because of its high convergence speed and good separation performance.

To solve the scale ambiguity, the back-projection [29] is applied, where the scale (which means the amplitude and phase at each frequency in the case of frequency-domain BSS) of the separated source is determined by the observed mixture at the selected microphone. In the distributed microphone array, the selection of the microphone is important because the input SNR varies considerably among the microphones. To obtain the result of back-projection to the closest microphone for each source in a blind manner, all possible results of back-projection are first calculated. Let $y_{ij}(t)$ be the ith separated source projected to the jth microphone. Then, we select $y_{i\sigma(i)}(t)$ such that $\sum_i \sum_t y^2_{i\sigma(i)}(t)$ is maximized, where $\sigma(i)$ represents a permutation of the index i.

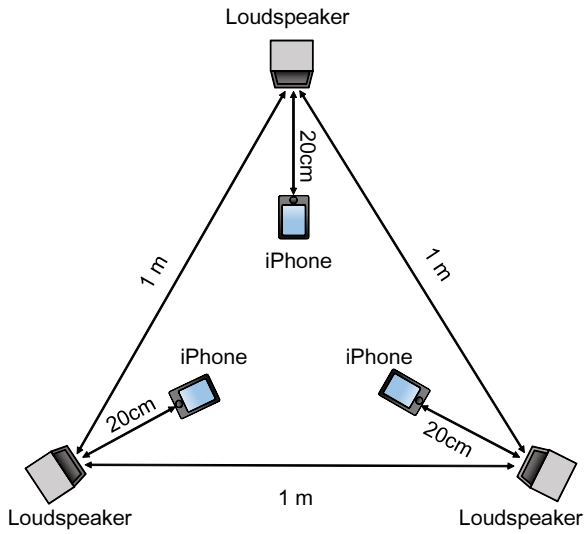

Figure 3: *Experimental setup*

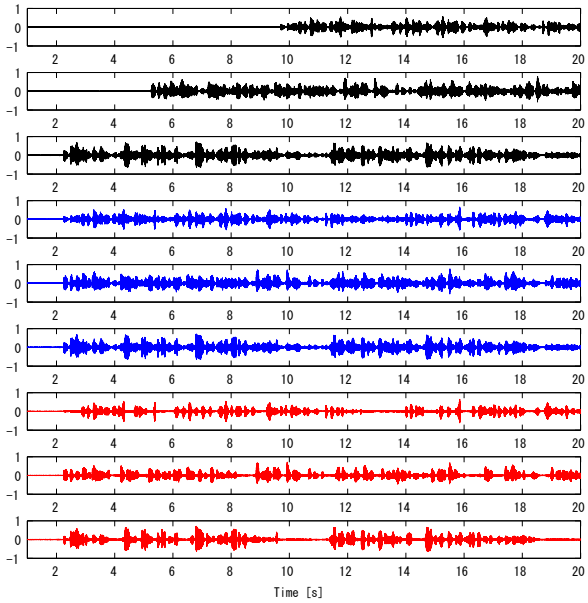

Figure 4: *Waveforms of observed (Mix), synchronized and separated signals (IVA w sync), respectively*

3.3. Automatic Speech Recognition System

The separated speech is analyzed and transcribed into text data by Julius version 4.3.1 [30]. As acoustic features, 12-dimensional mel-frequency cepstral coefficients (MFCCs), their deltas, and the delta power (25 dimensions in total) are calculated with a frame length of 25 ms and a frame shift of 10 ms, and cepstral mean normalization (CMN) is applied. The acoustic model was a GMM-HMM triphone, which was trained by using the ASJ-JNAS corpus (86 hours) [31].

4. Experimental Evaluation

4.1. Experimental Configuration

An experiment to evaluate the proposed system by speech recognition was conducted in an office room. Three loudspeakers (BOSE Computer MusicMonitor) connected to a PC were used to reproduce speech material simultaneously. The speech materials were broadcast news articles included in the Real World Computing Partnership (RWCP) SP99 Corpus (in Japanese). Each loudspeaker played back a set of 18-20 sentences spoken by a different speaker. The total length of the speech was 6 minutes.

Three mobile devices (iPhone 5 Model A1453) were placed at a distance of 20 cm from the loudspeakers. The distance between the loudspeakers was 1 m. Fig. 3 shows the setup of the experiment. The speech signals were individually recorded by the three iPhones. The sampling frequency was set to 16 kHz. After the recording had finished, the sound recorded at each device was uploaded to Dropbox storage using a Wi-Fi channel. The complete upload took less than 2 mintes and 25 seconds.

4.2. Evaluation Method

The ASR performance for the following signals was evaluated by the mora and word error rates.

- Mix: Unprocessed signals (recorded mixture).
- IVA w/o sync: Separated signals obtained by applying AuxIVA to the unprocessed signals.
- IVA w shift: Separated signals obtained by applying AuxIVA to the synchronized signals by only compensating the different time offsets (the sampling frequency mismatch remains).
- BM w shift: Separated signals obtained by applying a binary mask to the synchronized signals by only compensating the different time offsets (the sampling frequency mismatch remains).
- IVA w sync: Separated signals obtained by applying AuxIVA to the synchronized signals by compensating both the different time offsets and the sampling frequency mismatch.
- Ideal: Recorded signals when only one source was played (the ground truth of BSS).

Both of the blind synchronization and the blind source separation used here work in the STFT domain, but there are different trade-offs for determining the frame length. In this experiment, they were selected experimentally. In the blind synchronization, the frame length was 256 ms and the frame shift was the half of the frame. In BSS based on AuxIVA, the frame length was 1024 ms and the frame shift was the one fourth of the frame. The number of iterations in AuxIVA was 30.

For comparison, we applied a simple binary mask as an alternative BSS method (BM w shift). It was chosen because many other source separation methods do not work well on asynchronous signals. In our scenario, it was expected that each microphone would capture the closest source with the maximum input SNR. Thus, the binary mask was designed as follows:

$$M_i(t,\omega) = \begin{cases} 1 & (|X_i(t,\omega)|^2 \geq \sum_{j\neq i} |X_j(t,\omega)|^2) \\ 0 & (|X_i(t,\omega)|^2 < \sum_{j\neq i} |X_j(t,\omega)|^2) \end{cases},$$

where $X_i(t,\omega)$ denotes the signal recorded by ith microphone in the STFT domain. The separated signal was obtained by the inverse STFT of $M_i(t,\omega)X_i(t,\omega)$.

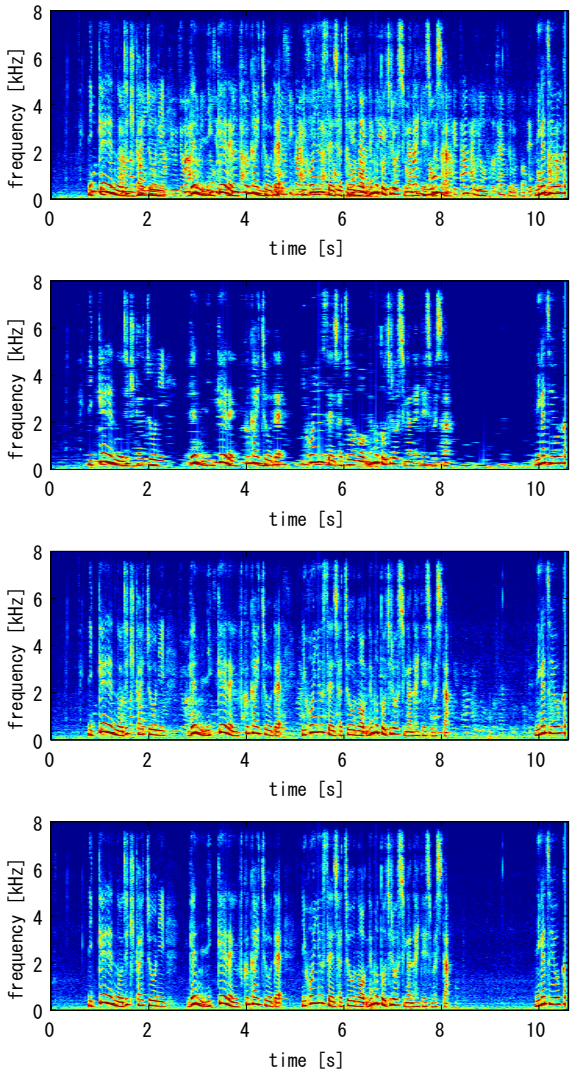

Figure 5: *Spectrograms of Mix, BM w shift, IVA w sync, and Ideal signals from top to bottom, respectively*

4.3. Experimental Results

Fig. 4 shows the waveforms of observed (Mix), synchronized, and separated signals (IVA w sync), respectively. The estimated sampling frequency mismatches between two channels were about 8.20 and -1.53 ppm. The estimated time offsets were -119188 and -48210 samples. The Fig. 5 shows the spectrograms of Mix, BM w shift, IVA w sync, Ideal signals, respectively.

The mora error rate and word error rate are shown in Fig. 6.The results of Mix had more substitution and injection errors because the other speaker's voice was also transcribed by ASR. Fig. 6 also shows that AuxIVA does not improve the ASR performance not only without synchronization but also with only compensation of the time offset. This indicates that compensating only the time offset is not sufficient for AuxIVA even it is small (about 8 ppm in this experiment). The binary mask slightly reduced the error rate but it is not sufficient. It is considered that the binary mask removed spectral information which was necessary for ASR. On the other hand, AuxIVA with the compensation of both the different time offsets and the sampling frequency mismatch showed a better error rates, giving results are much closer to those for Ideal. This indicates that BSS based on AuxIVA in this case successfully reduced the interference speech. The effect can be also confirmed in the waveforms shown in Fig. 4.

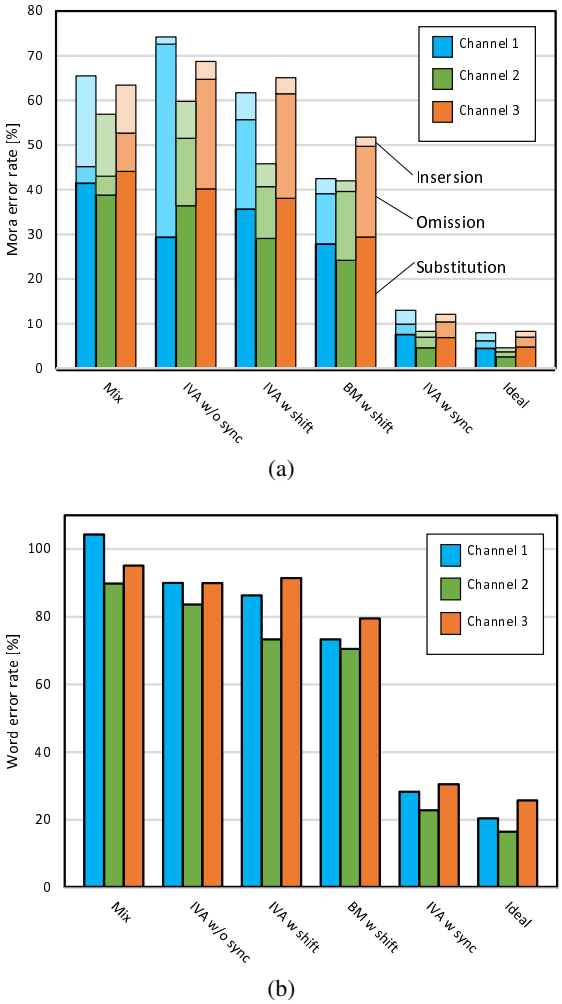

Figure 6: *The ASR evaluation results with (a) mora error rate and (b) word error rate*

5. Conclusion

In this paper, we presented a multi-talker speech recognition system based on blind source separation using iPhones and Dropbox. The blind synchronization technique based on the stationary spatial model well compensated both the difference in the time offset and the sampling frequency mismatch. The experimental results showed that the combination of blind synchronization and blind source separation contributed to significantly improving the ASR performance in the multiple speaker environment.

6. Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (B) (Grant Number: 25280069) and a Grant-in-Aid for Scientific Research (A) (Grant Number: 16H01735).

7. References

[1] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, A. Janin, M. Magimai-Doss, and J. Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system," *Multimodal Technologies for Perception of Humans* pp. 450–463, Springer Berlin Heidelberg, 2008.

[2] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. E. Hannani, and V. Wan, "Transcribing meetings with the AMIDA systems", *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 20 no. 2, pp. 486–498, 2012.

[3] G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tür, J. Dowding, and C. Frederickson, "The CALO meeting assistant system," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 18 no. 6, pp. 1601–1611, 2010.

[4] E. Shriberg, A. Stolcke and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," *Proc. INTERSPEECH*, pp. 1359–1362, 2001.

[5] R. Yokoyama, Y. Nasu, K. Iwano, and K. Shinoda, "Detection of overlapped speech using lapel microphones in meeting," *Speech Commun.*, vol. 55, pp. 941–949, 2013.

[6] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[7] D. C. Moore and L. A. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," *Proc. ICASSP*. vol. 5, pp. 497–500, 2003.

[8] D. Marino and T. Hain. "An analysis of automatic speech recognition with multiple microphones," *Proc. INTERSPEECH*, pp. 1281–1284, 2011.

[9] H. K. Maganti, D. Gatica-Perez and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 15, no. 8, pp. 2257–2269, 2007.

[10] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, and T. Nakatani, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 20 no. 2, pp. 499–513, 2012.

[11] W. Li, J. Dines, and H. Bourlard, "Non-linear mapping for multi-channel speech separation and robust overlapping spech recognition," *Proc. ICASSP*, pp. 3921–3924, 2009.

[12] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks - Part I: sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, pp. 5277–5291, 2010.

[13] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks - Part II: simultaneous & asynchronous node updating," *IEEE Trans. Signal Process.*, vol. 58, pp. 5292–5306, 2010.

[14] A. Bertrand, S. Doclo, S. Gannot, N. Ono and T. van Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Process.*, vol. 107, pp. 1-3, Feb. 2015.

[15] I. Himawan, I. McCowan and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Aud. Speech. Lang. Process.*, vol. 19, no. 4, pp. 661–676, 2011.

[16] S. Markovich-Golan, A. Bertrand, M. Moonen and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process.*, vol. 107, pp. 4–20, 2015.

[17] V. M. Tavakoli, J. R. Jensen, M. G. Christensen and J. Benesty, "Pseudo-coherence-based MVDR beamformer for speech enhancement with ad hoc microphone arrays" *Proc. ICASSP*, pp. 2659–2663, 2015.

[18] J. P. Dmochowski, Z. Liu and P. Chou, "Blind source separation in a distributed microphone meeting environment for improved teleconferencing," *Proc. ICASSP*, pp. 89–92, 2008.

[19] T. Ono, S. Miyabe, N. Ono and S. Sagayama, "Blind source separation with distributed microphone pairs using permutation correction by intra-pair TDOA clustering," *Proc. IWAENC*, Aug. 2010.

[20] K. Kinoshita and T. Nakatani, "Modeling inter-node acoustic dependencies with restricted Boltzmann machine for distributed microphone array based BSS," *Proc. ICASSP*, pp. 464–468, 2015.

[21] S. Miyabe, N. Ono and S. Makino, "Blind compensation of interchannel sampling frequency mismatch with maximum likelihood estimation in STFT domain," *Proc. ICASSP*, pp. 674–678, 2013,

[22] S. Miyabe, N. Ono and S. Makino, "Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording," *Proc. WASPAA*, Oct. 2013.

[23] S. Miyabe, N. Ono and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185–196, 2015.

[24] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *Proc. WASPAA*, pp. 189–192, 2011.

[25] R. Lienhart, I. Kozintsev, S. Wehr and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," *Proc. ICASSP*, pp. 840–843, 2003.

[26] E. Robledo-Arnuncio, T. S. Wada and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Proc. WASPAA*, pp. 34–37, Oct. 2007.

[27] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, 2008.

[28] T. Ono, S. Miyabe, N. Ono and S. Sagayama, "Blind source separation with distributed microphone pairs using permutation correction by intra-pair TDOA clustering," *Proc. IWAENC*, Aug. 2010.

[29] N. Murata, S. Ikeda and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[30] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," *Proc. APSIPA*, pp. 131–137, 2009

[31] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn.*, vol. 20, no. 3, pp. 199–206, 1999.