Spatial filter calibration based on minimization of modified LSD

Nobuaki Tanaka¹, Tetsuji Ogawa², Tetsunori Kobayashi¹

¹Dept. of Computer Science, Waseda University
²Waseda Institute for Advanced Study

Abstract

A new sound source separation method has been developed that is robust against individual variability in microphones and acoustic lines. A specific area that has a target sound source was enhanced by using a spatial filter developed by time-frequency masking. However, there is a strong likelihood that the spatial filters will be distorted due to the impact of individual variability in microphone characteristics and acoustic lines. To solve this problem, calibration of these spatial filters’ shapes was attempted using a modified log-spectral distance (MLSD) minimization criterion, which uses utterances made by each individual (i.e., a sound source) at the desired positions. The effectiveness of this spatial filter calibration was experimentally verified in speech recognition experiments; MLSD-based calibration had fewer word errors than the cases without calibration and calibration using other criteria.

Index Terms: sound source separation, time-frequency masking, spatial filter calibration, modified LSD

1. Introduction

To achieve high-performance speech recognition systems in noisy environments, a desired sound should be precisely separated from noisy speech. Adaptive beamforming and time-frequency masking are frequently applied to sound source separation using microphone arrays. Adaptive beamforming techniques such as an independent component analysis (ICA) [1] sequentially estimate direction-of-arrivals (DOAs). Therefore, these methods induce unavoidable delays for converging filters to deal with moving sources. In contrast, we have proposed area-enhancement-based sound source separation [3, 4, 5], which enhances not just a sound source but an area that includes a target sound source by using spatial filters developed by time-frequency masking [2]. This method is robust against sound source movements in the enhanced area. The spatial filters used, however, will be distorted due to impacts of individual variability in microphone characteristics and acoustic lines [5]. The distortion of the spatial filters negatively affects performance of sound source separation systems, especially when a spatial filter with a sharp directivity must be developed because the target and disturbance sources are closely located.

This paper proposes a new method for calibrating the spatial filters to eliminate the impact of their distortion and obtain desired shapes of the spatial filters. The proposed method assumes that each individual (i.e., each sound source) makes an utterance at the desired position; the utterance is recorded and used for reference speech. Then, the spatial filters are optimized to minimize the spectral distance between the reference sound and the target sound estimated using sound source separation. In this study, we propose a modified log-spectral distance (MLSD) and verify the effectiveness of the MLSD for spatial filter calibration criterion.

The rest of this paper is organized as follows. Section 2 briefly reviews area-enhancement-based sound source separation using time-frequency masking. Section 3 presents the proposed spatial filter calibration method. Section 4 defines the MLSD in detail. Section 5 evaluates the proposed method in speech recognition experiments. Finally, Section 6 presents some concluding remarks.

2. Area-enhancement-based sound source separation

We assume a situation where there are two sound sources and two omnidirectional microphones as shown in Fig. 1(a). We regard one sound source as a target and the other as a disturbance. Let \( x_1(t) \) and \( x_2(t) \) be signals received by the omnidirectional microphones Ch1 and Ch2 respectively shown in Fig. 1(a) at a discrete time \( t \). \( X_1(\omega, k) \) denotes an STFT coefficient of \( x_1(t) \), where \( k \) and \( \omega \) denote a discrete frame and a discrete frequency, respectively.

To separate the target speech from observed signals, we developed a spatial filter that enhances a sound source in the colored area shown in Fig. 1(b). To develop the spatial filter, two cardioid-shaped beamformers were developed and then time-frequency masking was carried out using outputs of these beamformers. \( C_1(\omega, k) \) and \( C_2(\omega, k) \) denote spectral components of the outputs of these cardioid-shaped beamformers developed by delay addition followed by subtraction using \( X_1(\omega, k) \) and \( X_2(\omega, k) \). \( C_1(\omega, k) \) and \( C_2(\omega, k) \) are computed as follows:

\[
\begin{align*}
C_1(\omega, k) &= X_1(\omega, k) - \exp(-j\omega \tau_1) \cdot X_2(\omega, k) \quad (1) \\
C_2(\omega, k) &= X_2(\omega, k) - \exp(-j\omega \tau_2) \cdot X_1(\omega, k), \quad (2)
\end{align*}
\]

where \( \tau_1 \) and \( \tau_2 \) denote delays corresponding to microphone spacings. Directivity patterns of \( C_1 \) and \( C_2 \) are shown in

![Figure 1: Development of spatial filter](image-url)
In this section, we propose a spatial filter calibration method. Figure 2 shows the block diagram of the proposed method. The calibration is carried out as follows:

1. An utterance spoken by a target speaker is picked up by a microphone array. Here, T denotes the picked up utterance.
2. An utterance spoken by a disturbance speaker is picked up by the same microphone array that picked up the target speaker’s voice. D denotes the picked up utterance. Note that T and D should be independently recorded and not be mixed.
3. T is superposed with D to obtain T + D.
4. The target speaker’s utterance is estimated from T + D by sound source separation with the parameter Φ. In this case, TΦ denotes the estimate of the target speaker’s utterance when using Φ.
5. Similarity (e.g., spectral distance) between T and TΦ is measured.

6. Steps 4 and 5 are iterated until Φ is converged; the distance is minimized.

However, it is difficult to analytically compute optimal Φ. Therefore, we compute Φ using the grid search-based approach described in 5.2.3. By using this method, we can reduce the impact of the distortion in spatial filters.

4. Modified LSD

In this section, we define the criterion for calibrating the spatial filters. In this study, the spectral distance between T and TΦ was applied to the criterion. A log-spectral distance (LSD) is frequently used for evaluating the similarity between the clean speech and the separated speech [6, 7]. The LSD is computed as follows:

$$\text{LSD (dB)} = \frac{1}{W K} \sum_{\omega,k} 10 \log_{10} \left\{ \frac{|\hat{S}(\omega,k)|}{|S_{ref}(\omega,k)|} \right\}^2$$

where $\hat{S}(\omega,k)$ and $S_{ref}(\omega,k)$ denote spectral components of the estimated target speech, which is given by sound source separation with Φ, and that of the reference sound recorded as described in Section 3, respectively; and K and W, the number of discrete frames and that of discrete frequencies, respectively.

We attempted to improve performance of recognizing speech after sound source separation. Since the preliminary experiments showed that the LSD did not highly correlate with speech recognition system performance, we propose a new criterion named “modified log-spectral distance (MLSD)” for spatial filter calibration. MLSD is computed as follows:

$$\text{MLSD}_{(z)} (dB) = \frac{1}{W K} \sum_{\omega,k} \left\{ \mu_{\omega,k} \cdot 10 \log_{10} \left\{ \frac{|\hat{S}(\omega,k)|}{|S_{ref}(\omega,k)|} \right\} \right\}^2$$

where $\mu_{\omega,k}$ is computed as follows:

$$\mu_{\omega,k} = \begin{cases} z, & \text{if } |\hat{S}(\omega,k)| > |S_{ref}(\omega,k)| \\ 1, & \text{otherwise} \end{cases}$$

where $z$ denotes a tuning parameter of the MLSD. $z$ should be $z > 1$. If $z = 1$, the MLSD is equal to the LSD. The MLSD is designed on the basis of heuristics in recognizing the speech after noise reduction processing; it is more important to suppress the noise clearly (even though the distortion increases) than to reduce the distortion (even though the noise increases) in target speech. When computing the MLSD, spectral components where power spectra of the estimated sound are larger than those of the reference sound are weighted. In other words, the MLSD decreases as estimated spectral components not appearing in the reference sound increase, unlike the LSD. Applying the MLSD to the criterion for spatial filter calibration could be useful for improving performance of recognizing speech after sound source separation with Φ.

5. Speech recognition experiment

We carried out speech recognition experiments to verify the effectiveness of the proposed spatial filter calibration method. The proposed method was evaluated in terms of speech recognition system performance on the basis of word accuracy.

---

Figure 2: Block diagram of proposed calibration method
5.1. Evaluation items

Experimental comparison was carried out from the following viewpoints:

- Effects of spatial filter calibration on performance of speech recognition systems
- Effectiveness of calibration criteria

In addition to the LSD and MLSD, we applied a cepstral distance (CD) to an objective function for calibration. The CD is computed as follows:

\[
CD \ (\text{dB}) = \frac{1}{K} \sum_{k} \frac{10}{\ln 10} \left\{ \frac{2}{D} \sum_{l=1}^{D} \left| \hat{c}_{k,l} - c_{k,l} \right|^2 \right\}, \tag{9}
\]

where \( D \) denotes the dimensionality of cepstral coefficients (\( D = 12 \) in this paper); \( \hat{c}_{k,l} \) and \( c_{k,l} \) denote \( l \)-th cepstral coefficients at discrete frame \( k \) of the estimated target speech and that of the reference speech recorded as described in Section 3, respectively.

5.2. Experimental condition

5.2.1. Speech materials

The target speech utterances consisted of 100 sentences, which were spoken by 23 male speakers; these sentences were taken from a continuous speech database that contained sentences from Japanese newspaper articles [8]. As for disturbance speech utterances, 100 sentences were selected from the same database; however, these sentences were different from the target speech utterances. Each disturbance speech utterance lasted approximately as long as the corresponding target speech utterance.

For calibration, we used one target speech utterance and one disturbance speech utterance from the aforementioned 100 sentences, and the remaining 99 utterances for evaluating speech recognition systems.

5.2.2. Speech recording

Figure 3 shows the recording environment. We placed a target sound source in one of seven directions \( \theta_{\text{tar}} = 0^\circ, \pm 30^\circ, \pm 60^\circ \), or \( \pm 90^\circ \) at 1.0 m from the microphone array and a disturbance sound source in one of seven directions \( \theta_{\text{dist}} = 0^\circ, \pm 30^\circ, \pm 60^\circ \), or \( \pm 90^\circ \) at 1.0 m from the microphone array. The target and disturbance speech utterances were recorded separately. The 100 utterances were played back through a loudspeaker placed at each position and then recorded with the microphones. Both target and disturbance speech had 700 utterances in total.

Table 1: Conditions for grid search. "(prev)" denotes parameter estimated at previous iteration step.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial value</th>
<th>Search range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>( (\theta_{\text{tar}} + \theta_{\text{dist}})/2 )</td>
<td>( \theta^{\text{(prev)}} \pm 30^\circ )</td>
</tr>
<tr>
<td>( \tau_1 )</td>
<td>0.03/340(s)</td>
<td>( \tau_1^{\text{(prev)}} \pm 0.02/340 )</td>
</tr>
<tr>
<td>( \tau_2 )</td>
<td>0.03/340(s)</td>
<td>( \tau_2^{\text{(prev)}} \pm 0.02/340 )</td>
</tr>
<tr>
<td>( u )</td>
<td>0.5</td>
<td>( u^{\text{(prev)}} \pm 0.25 )</td>
</tr>
</tbody>
</table>

For the sound source separation experiment, a disturbance speech utterance was electronically superposed on the corresponding target speech utterance at an SNR of 0 dB to simulate simultaneous speech.

5.2.3. Spatial filter calibration

The spatial filter was developed as described in Section 2. In this case, beamforming and time-frequency masking computed as Eqs. 1–3 were modified as follows:

\[
C_1'(\omega, k) = u \cdot X_1(\omega, k) - \exp(-j\omega \tau_1) \cdot (1 - u) \cdot X_2(\omega, k) \tag{10}
\]

\[
C_2'(\omega, k) = (1 - u) \cdot X_2(\omega, k) - \exp(-j\omega \tau_2) \cdot u \cdot X_1(\omega, k) \tag{11}
\]

\[
T'(\omega, k) = \begin{cases} X_1(\omega, k), & \text{if } |C_1'(\omega, k)| > \alpha_{\omega, \theta} |C_2'(\omega, k)| \\ \beta \cdot X_1(\omega, k), & \text{otherwise} \end{cases} \tag{12}
\]

where \( u \) is a parameter controlling the ratio of gains between microphones Ch1 and Ch2. If \( u = 0.5 \), the output of time-frequency masking in Eq. 12 equals that in Eq. 3. We attempt to optimize parameters \( \Phi = \{ \theta, \tau_1, \tau_2, u \} \). The optimization of \( \theta \) in Fig. 1(b) indicates that the effect of misalignment in microphone arrangement is reduced. The optimization of \( \tau_1 \) and \( \tau_2 \) indicates that the difference in phase characteristics of microphones is compensated for. The optimization of \( u \) indicates that the difference in amplitude characteristics of microphones is compensated for too. These parameters were estimated by using the following grid search algorithms:

1. Search \( \theta \) optimizing the calibration criterion (e.g., minimizing the MLSD between \( \mathbf{T} \) and \( \mathbf{T}_p \) ) using grid search.
2. Search \( \tau_1 \) optimizing the calibration criterion using grid search.
3. Search \( \tau_2 \) optimizing the calibration criterion using grid search.
4. Search \( u \) optimizing the calibration criterion using grid search.
5. Steps 1–4 were iterated until no parameter could be updated or the number of iterations reached ten.

Table 1 lists conditions for grid search, such as conditions for initialization of parameters and search ranges in grid search.

5.2.4. Sound source separation and speech recognition

Experimental conditions for speech similarity computation, sound source separation, and acoustic feature extraction are listed in Tables 2, 3, and 4, respectively.
Table 2: Conditions for speech similarity computation

<table>
<thead>
<tr>
<th>sampling frequency</th>
<th>32 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame length</td>
<td>64 ms</td>
</tr>
<tr>
<td>frame shift</td>
<td>32 ms</td>
</tr>
<tr>
<td>analysis window</td>
<td>Hanning window</td>
</tr>
<tr>
<td>analysis range</td>
<td>300–5500 Hz</td>
</tr>
</tbody>
</table>

Table 3: Conditions for sound source separation

<table>
<thead>
<tr>
<th>sampling frequency</th>
<th>32 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame length</td>
<td>64 ms</td>
</tr>
<tr>
<td>frame shift</td>
<td>16 ms</td>
</tr>
<tr>
<td>analysis window</td>
<td>Hanning window</td>
</tr>
<tr>
<td>analysis range</td>
<td>300–5500 Hz</td>
</tr>
</tbody>
</table>

Table 4: Conditions for speech recognition

<table>
<thead>
<tr>
<th>sampling frequency</th>
<th>16 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame length</td>
<td>25 ms</td>
</tr>
<tr>
<td>frame shift</td>
<td>10 ms</td>
</tr>
<tr>
<td>analysis window</td>
<td>Hamming window</td>
</tr>
<tr>
<td>feature parameters</td>
<td>12 MFCCs, 12 ΔMFCCs, and a Δlog energy</td>
</tr>
</tbody>
</table>

To reduce effects of spatial aliasing in beamformers, we used a band-pass filter of 300–5500 Hz. The flooring coefficient used in time-frequency masking (i.e., $\beta$ in Eq. 12) was 0.01.

For speech recognition systems, speech data after sound source separation were down sampled from 32 kHz to 16 kHz.

Acoustic models were trained with 20414 sentences spoken by 133 male speakers, taken from the ASJ database [8], which consisted of Japanese newspaper article sentences (ASJ-JNAS) and phonetically balanced sentences (ASJ-PB) recorded with close-talking microphones. We used tied-state triphones with 2000 states. The distribution function in each state of the models was represented by a 16-mixture Gaussian distribution with diagonal covariances.

We used word trigram language models that were constructed using a lexicon with a 20 K-word vocabulary.

5.3. Experimental results

Table 5 lists word accuracies for various recording conditions. In the case of “w/o calibration,” the initial parameters in Table 1 were used for sound source separation. In this experiment, the LSD, CD and MLSD were applied to objective functions for spatial filter calibration criteria. As for the MLSD, $z$ in Eq. 8 was 1.5 and 2.0. In this table, the highest accuracy for each condition is underlined and written in bold. In most situations, proposed MLSD-based calibration performance was best. Without any calibration, word accuracies were not symmetric in source positions due to the effect of individual variability in the microphone characteristics or acoustic lines. In addition, using the MLSD improved the word accuracy more than that of the LSD or the CD. This result indicates that the MLSD might have higher correlation with the performance of speech recognition systems than the other criteria. In the case of $(\theta_{src}, \theta_{dst}) = (-90^\circ, +90^\circ)$, MLSD-based calibration with $z = 2.0$ exhibited approximately 5% higher accuracy than the case “w/o calibration.” Therefore, the proposed calibration method can improve performance even when the spacing between sound sources is large (i.e., high performance is achieved without calibration).

6. Conclusion

We proposed the spatial filter calibration method for area-enhancement-based sound source separation and the MLSD minimization criterion for optimizing spatial filters. We verified that the method improved the performance of recognizing separated speech data more than that without any calibration and that using the other criteria-based calibration.

7. References