Multiple Sound Source Localization Based on Interchannel Phase Differences in All Frequencies with Spectral Masks

Hyungchan Song and Jong Won Shin

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
{shchan420, jwshin}@gist.ac.kr

Abstract

One of the most widely used cues for sound source localization is the interchannel phase differences (IPDs) in the frequency domain. However, the spatial aliasing makes the utilization of the IPDs in the high frequencies difficult, especially when the distance between the microphones is high. Recently, the phase replication method which considers the direction-of-arrival (DoA) candidates corresponding to all the possible unwrapped phase differences in all frequency bins was proposed. However, high frequency bins with possible spatial aliasing contribute more when constructing initial DoA histograms compared with low frequency bins, which may not be desirable for source localization. In this paper, we propose to utilize the IPDs in all the frequency bins with equal weights regardless of maximum number of phase wrapping in that frequency for dual microphone sound source localization. We applied spectral masks based on local signal-to-noise ratios and coherences between microphone signals to exclude time-frequency bins without directional audio signal from the DoA histogram construction. Experimental results show that the proposed method results in more distinct peaks in the DoA histogram and outperforms the conventional method in various noisy and reverberant environments.

Index Terms: multiple sound source localization, interchannel phase difference, spatial aliasing

1. Introduction

Sound source localization (SSL) plays a crucial role in many of the multichannel speech processing such as speech enhancement and recognition [1–13]. One of the most widely used spatial cues of many SSL algorithms in the frequency domain is interchannel phase difference (IPD) [2–13]. However, spatial aliasing makes it difficult to utilize the IPDs in high frequencies for direction-of-arrival (DoA) estimation. Many SSL algorithms [2–8] assume that the minimum distance between microphones is short enough so that the whole frequency bins are free from spatial aliasing, which is not achievable for some of the devices such as modern smartphones. On the other hand, several SSL algorithms [9–12] try to unwrap the phase differences to overcome spatial aliasing problem, with the assumption that the number of microphones is greater than the number of sources. Chen et al. [13] proposed a phase replication method considering all the possible unwrapped phase difference candidates in all frequency bins for multiple source localization. Based on theoretical analysis that the correct DoA estimates would be present in many frequency bins while the aliasing DoAs would be estimated only in a limited number of bins, the DoA histogram is constructed for which the peaks in the histogram after post-processing are regarded as DoA estimates.

In this paper, we propose to improve the phase replication method in [13] so that each frequency bin with directional audio signal contributes equally when constructing the DoA histogram. Specifically, we have applied spectral masks based on the estimated local signal-to-noise ratios (SNRs) and the coherence between two microphone signals to rule out the T-F bins with diffuse noise only. Each of the survived frequency bins contributes equally to the construction of the DoA histogram by probabilistic voting to the possible DoA candidates corresponding to the IPDs in that frequency. Experimental results show the proposed SSL outperformed the conventional method in terms of the minimum root mean square error in various noisy and reverberant environments.

2. Spatial aliasing and phase replication method

Let $X_1(n, k)$ and $X_2(n, k)$ denote the short-time Fourier transform (STFT) coefficients of the primary and secondary microphone signals for the $k$-th frequency bin in the frame $n$, respectively. The observed IPD $\Delta \phi(n, k)$ is defined as the difference between phases of $X_1(n, k)$ and $X_2(n, k)$:

$$\Delta \phi(n, k) = \angle \{X_1(n, k)X_2^*(n, k)\}. \quad (1)$$

Under the far-field assumption, the relationship between IPD $\Delta \phi(n, k)$ and the DoA $\theta$ with respect to the broadside direction is given as

$$\Delta \phi(n, k) = \frac{2\pi f_k \sin \theta/d}{c}, \quad (2)$$

where $d$ denotes the distance between microphones, $c$ is the speed of sound, and $f_k$ represents the center frequency for the $k$-th frequency bin. $\Delta \phi$ in the equation (2) can be outside the principal range, $[-\pi, \pi]$, for high frequencies, which results in the wrapping of the phase, i.e., $\Delta \phi = \Delta \phi - 2\pi l$ for an integer $l$. The phenomenon that multiple DoA candidates exist for a given observed IPD due to the phase wrapping is called spatial aliasing. The lowest frequency affected by the spatial aliasing is dependent on the DoA $\theta$ and the distance between microphones $d$. Given $d$, the lowest frequency that may be affected by spatial aliasing for the signal coming from any DoA, $f_{a0}$, becomes

$$f_{a0} = \min_{\theta} \frac{c}{2|\sin \theta/d} = \frac{c}{2d} \quad (3)$$

and the STFT bin index corresponding to $f_{a0} = c/2d$ is denoted as $N_{a0}$.

Although spatial aliasing makes the use of the IPDs in high frequencies more complicated, the IPDs in the high frequencies still have useful information on the DoAs of the input signals. Chen et al. [13] proposed a phase replication method considering all the possible unwrapped phase difference candidates in
Figure 1: Block diagram of the proposed multiple sound source localization system.

all frequency bins. As the first step of this method, a replicated IPD matrix \( \Pi(n) \in \mathbb{R}^{(2AK+1)\times K} \) is constructed as follows:

\[
\Pi(n) = \begin{bmatrix}
0 & \cdots & 0 & \cdots & \Delta \phi(n, K) - 2\pi A_k \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & \Delta \phi(n, N_{\text{min}}) - 2\pi & \cdots & \Delta \phi(n, K) - 2\pi \\
\Delta \phi(n, 1) & \cdots & \Delta \phi(n, N_{\text{min}}) & \cdots & \Delta \phi(n, K) \\
0 & \cdots & \Delta \phi(n, N_{\text{min}}) + 2\pi & \cdots & \Delta \phi(n, K) + 2\pi \\
0 & \cdots & 0 & \cdots & \Delta \phi(n, K) + 2\pi A_k \\
\end{bmatrix},
\]

where \( K \) denotes the number of frequency bins, and \( A_k \) is the maximum number of unwrapping \( l \) for the \( k \)-th frequency bin, which is given as

\[
A_k = \left\lfloor \frac{k - N_{\text{min}}}{2N_{\text{min}}} \right\rfloor.
\]

It is noted that the “\( \emptyset \)” entries are just place holders that will be omitted when constructing the DoA histogram. The replicated IPD matrix \( \Pi(n) \) can be easily converted into a DoA matrix \( \Theta(n) \), for which the \( k \)-th column \( \Theta_k(n) \) becomes

\[
\Theta_k(n) = \arcsin \left( \frac{c \cdot \Pi_{k}(n)}{2\pi f_s d} \right),
\]

where \( \Pi_k(n) \) denotes the \( k \)-th column of the replicated IPD matrix \( \Pi(n) \). The DoA histogram \( h_k \) for the frequency \( k \) is constructed by counting the number of entries in \( \Theta_k(n) \) that falls in the specific ranges of DoA for all frames. Then, the utterance-level DoA histogram \( h \) is given by

\[
h = \sum_{k=1}^{K} h_k,
\]

in which the peaks are regarded as the estimates for the DoAs. The theoretical background is that many T-F bins will produce the DoA estimates at the true DoAs, while the DoA estimates due to the spatial aliasing may differ for each frequency.

Furthermore, two post-processing are proposed in [13]. After finding more number of peaks than the number of sources in the initial histogram \( h \), a subvector selection method is applied to select one DoA candidate for each frequency for each of the peaks, assuming that the peak corresponds to one of the true DoAs and thus the correct phase difference unwrapping is known. By taking the maximum among the modified histograms of the selected subvectors after applying soft masks depending on the DoAs, a modified histogram is constructed. Final DoA estimates are the peaks in the modified histogram after some smoothing.

3. Proposed method

In this paper, we propose two modifications on the method in [13] for more robust SSL. The block diagram of the proposed method is shown in the Figure 1. Firstly, we propose to apply spectral masks to exclude the T-F bins for which the IPDs are corrupted by the background noises and interfering sources severely, as in [11]. To use the frequency bins with high local SNRs only, we estimate the SNR-based mask \( M_{\text{SNR}}(n, k) \) as

\[
M_{\text{SNR}}(n, k) = \begin{cases} 1, & \lambda(n, k) > \lambda_{TH} \\ 0, & \text{otherwise} \end{cases},
\]

in which \( \lambda_{TH} \) is a predefined threshold and

\[
\lambda(n, k) = \min \left( \frac{P_{m}(n, k)}{P_{s}(n, k)}, 1 \right),
\]

is the estimate of the local SNR where \( P_{m}(n, k) = |X_m(n, k)|^2 \) denote the power of the \( m \)-th microphone signal, and \( P_{s}(n, k) \) denotes the estimated power of the noise in the \( m \)-th microphone signal obtained by the method in [14]. We also use the coherence-based mask to only use the frequency bins where two microphone signals are coherent:

\[
M_{\text{coh}}(n, k) = \begin{cases} 1, & \gamma(n, k) > \gamma_{TH} \\ 0, & \text{otherwise} \end{cases},
\]

in which \( \gamma_{TH} \) is a predefined threshold, and

\[
\gamma(n, k) = \frac{E \left[ X_1(n, k) X_2^*(n, k) \right]}{E \left[ |X_1(n, k)|^2 \right] E \left[ |X_2(n, k)|^2 \right]},
\]

is the coherence, where the expectation \( E[\cdot] \) is approximated by a time average as

\[
E \left[ \alpha(n, k) \right] = \frac{1}{C+1} \sum_{n'=n-C}^{n} \alpha(n', k),
\]

in which \( C \) denotes the number of consecutive time frames. The final spectral mask \( M_{TF}(n, k) \) is constructed as a product of
If the next peak is lower than the threshold or the maximum number of sources \( P_N \) is reached.

4. Experimental results

The sound source localization performance was evaluated for the proposed and the conventional methods under various environments. To analyze the contributions of the proposed spectral masks and the probabilistic voting schemes for the performance improvement, we also investigated the performance with the spectral masks and conventional histogram construction. The room configurations and the locations of two microphones and two sound sources are illustrated in Figure 3. We simulated an office room of \( 6 \times 6 \times 3 \text{ m} \) using the image method [17]. Two microphones are located at the center of the room, 14 cm away from each other, which is a typical distance between microphones for modern smartphones. Spatial aliasing would be severer for this form factor, and thus it needs to be taken care of to localize the sound source in the frequency domain. The reverberation time RT60 ranged from 200 ms to 1 s with 200 ms intervals. The first sound source was located at the DoA of \(-30^\circ, 2 \text{ m}\) away from the center of two microphones. The second sound source was also \(2 \text{ m}\) away from the center of two microphones, and the DoAs were \(-10^\circ, 10^\circ, \) and \(30^\circ\). We selected 5 male and 5 female speech utterances from the TIMIT corpus [18] as the signals from the first sound source, while another 10 utterances from different 5 male and 5 female speakers were used as the signal from the second sound source. The powers of two source signals were set to be the same with each other. In addition, diffuse noise was generated by using the arbitrary noise field (ANF)-generator [19], in which the babble noise from the NOISEX-92 database [20] was employed. The diffuse noise was mixed with two sound sources with the SNRs from 0 dB to 20 dB at 5 dB intervals. The parameters were set as \( K = 257, \lambda_{TH} = 5 \text{ dB}, \gamma_{TH} = 0.9, \beta = 0.5, C = 5, \) and \( P_N = 2, \) and the histogram was constructed for the DoA bins with 5 degree intervals.

We compared \( \text{minRMSE} \) [8] for the proposed and the conventional methods [13]. As the number of sound sources...
Table 1: Average minRMSE’s for the conventional method ([13]), that with spectral masks only (mask), and the proposed method (Prop.) depending on the SNR and the DoA of source #2 when RT60 = 200 ms.

<table>
<thead>
<tr>
<th>SNR</th>
<th>method</th>
<th>source#2</th>
<th>-10°</th>
<th>10°</th>
<th>30°</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 dB</td>
<td>[13]</td>
<td>mask</td>
<td>17.60</td>
<td>9.11</td>
<td>1.85</td>
<td>7.71</td>
</tr>
<tr>
<td>15 dB</td>
<td>[13]</td>
<td>mask</td>
<td>11.29</td>
<td>6.67</td>
<td>2.03</td>
<td>7.00</td>
</tr>
<tr>
<td>20 dB</td>
<td>[13]</td>
<td>mask</td>
<td>8.87</td>
<td>5.23</td>
<td>2.11</td>
<td>6.79</td>
</tr>
</tbody>
</table>

Average: 10.56, 8.17, 12.59, 1.85, 9.46, 6.35, 15.60, 12.65, 8.01, 11.79, 9.85, 17.60, 5.49, 5.26, 9.48, 5.32.

Table 2: Average minRMSE’s for the conventional method ([13]), that with spectral masks only (mask), and the proposed method (Prop.) depending on the reverberation time and the DoA of source #2 when SNR = 20 dB.

<table>
<thead>
<tr>
<th>RT60</th>
<th>method</th>
<th>source#2</th>
<th>-10°</th>
<th>10°</th>
<th>30°</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 s</td>
<td>[13]</td>
<td>mask</td>
<td>22.63</td>
<td>15.60</td>
<td>13.96</td>
<td>10.52</td>
</tr>
<tr>
<td>0.8 s</td>
<td>[13]</td>
<td>mask</td>
<td>22.86</td>
<td>14.35</td>
<td>12.00</td>
<td>10.76</td>
</tr>
<tr>
<td>0.6 s</td>
<td>[13]</td>
<td>mask</td>
<td>22.98</td>
<td>11.64</td>
<td>10.49</td>
<td>10.56</td>
</tr>
<tr>
<td>0.4 s</td>
<td>[13]</td>
<td>mask</td>
<td>23.99</td>
<td>7.82</td>
<td>8.30</td>
<td>9.48</td>
</tr>
<tr>
<td>0.2 s</td>
<td>[13]</td>
<td>mask</td>
<td>12.14</td>
<td>5.23</td>
<td>2.11</td>
<td>6.35</td>
</tr>
</tbody>
</table>


5. Conclusions

In this paper, we propose a sound source localization method to improve the phase replication method so that each frequency bin with a directional audio signal contributes equally when constructing the DoA histogram. Spectral masks based on the local SNR and the coherence between microphone signals were constructed to exclude the T-F bins without directional audio signals. The probabilistic voting scheme was also applied so that high frequencies with spatial aliasing do not contribute more to the sound source localization than low frequencies. Experimental results showed that the proposed method outperformed the conventional approach under various noisy and reverberant environments.

6. Acknowledgements

This research was funded in part by the National Research Foundation of Korea grant number NRF-2019R1A2C2089324, Samsung System LSI (SLSI-201801GE002S), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).
References


