INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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 $X_2(n,k)$:

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

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

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

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. $\Delta\tilde{\phi}$ 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., $\Delta\phi = \Delta\tilde{\phi} - 2\pi l$ 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 θ 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

$$f_{a0} = \min_{\theta} \frac{c}{2|\sin\theta| d} = \frac{c}{2d}, \tag{3}$$

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

http://dx.doi.org/10.21437/Interspeech.2021-1178

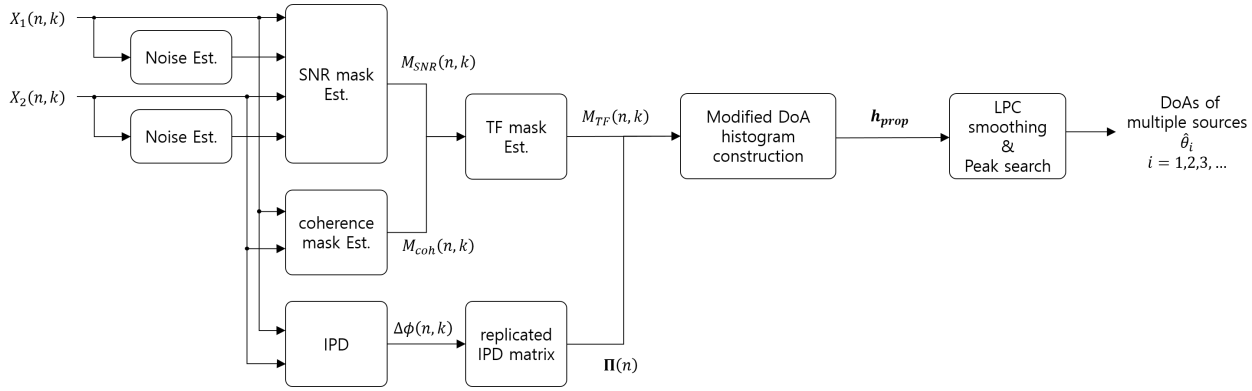


Figure 1: *Block diagram of the proposed multiple sound source localization system.*

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

$$\mathbf{\Pi}(n) = \begin{bmatrix} \emptyset & \cdots & \emptyset & \cdots & \Delta\phi(n,K) - 2\pi A_K \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \emptyset & \cdots & \Delta\phi(n,N_{a0}) - 2\pi & \cdots & \Delta\phi(n,K) - 2\pi \\ \Delta\phi(n,1) & \cdots & \Delta\phi(n,N_{a0}) & \cdots & \Delta\phi(n,K) \\ \emptyset & \cdots & \Delta\phi(n,N_{a0}) + 2\pi & \cdots & \Delta\phi(n,K) + 2\pi \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \emptyset & \cdots & \emptyset & \cdots & \Delta\phi(n,K) + 2\pi A_K \end{bmatrix}, \tag{4}$$

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

$$A_k = \lceil \frac{k - N_{a0}}{2N_{a0}} \rceil. \tag{5}$$

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

$$\mathbf{\Theta}_k(n) = \arcsin\left(\frac{c \cdot \mathbf{\Pi}_k(n)}{2\pi f_k d}\right), \tag{6}$$

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

$$\boldsymbol{h} = \sum_{k=1}^{K} \boldsymbol{h}_k, \tag{7}$$

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 $\boldsymbol{h}$, 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 histrogram 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 [1]. To use the frequency bins with high local SNRs only, we estimate the SNR-based mask $M_{SNR}(n,k)$ as

$$M_{SNR}(n,k) = \begin{cases} 1, & \lambda(n,k) > \lambda_{TH} \\ 0, & otherwise \end{cases}, \tag{8}$$

in which λ_{TH} is a predefined threshold and

$$\lambda(n,k) = \min\left(\frac{P_{x_1}(n,k)}{P_{v_1}(n,k)} - 1, \frac{P_{x_2}(n,k)}{P_{v_2}(n,k)} - 1\right), \tag{9}$$

is the estimate of the local SNR where $P_{x_m}(n,k) = |X_m(n,k)|^2$ denote the power of the m-th microphone signal, and $P_{v_m}(n,k)$ 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:

$$M_{coh}(n,k) = \begin{cases} 1, & \gamma(n,k) > \gamma_{TH} \\ 0, & otherwise \end{cases}, \tag{10}$$

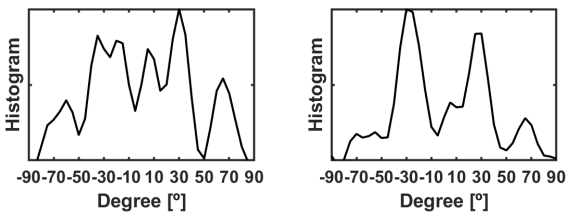
in which γ_{TH} is a predefined threshold, and

$$\gamma(n,k) = \left| \frac{E\left[X_1(n,k)X_2^*(n,k)\right]}{E\left[X_1(n,k)X_1^*(n,k)\right] E\left[X_2(n,k)X_2^*(n,k)\right]} \right| \tag{11}$$

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

$$E\left[\alpha(n,k)\right] = \frac{1}{C+1} \sum_{n'=n-C}^{n} \alpha(n',k), \tag{12}$$

in which C denotes the number of consecutive time frames. The final spectral mask $M_{TF}(n,k)$ is constructed as a product of



(a) (b)

Figure 2: *DoA histograms for (a) applying the spectral masks only, and (b) applying both the spectral masks and probabilistic voting when two sources are located at* -30° *and* 30°.

two masks:

$$M_{TF}(n,k) = M_{SNR}(n,k) \cdot M_{coh}(n,k). \quad (13)$$

Then, the DoA histogram for the masked DoA matrix in the frequency k, $\hat{\boldsymbol{h}}_k$, is constructed only with $\boldsymbol{\Theta}_k(n)$ for which $M_{TF}(n,k)$ is 1.

In the construction of the initial histogram $\boldsymbol{h}$ in (7), the k-th frequency bin provides $2A_k + 1$ candidate DoAs, which means that the high frequencies suggests more number of DoA candidates than the low frequencies. Although theoretical analysis is provided in [13] for clean scenarios, it may not be desirable for high frequencies to contribute more for the histogram construction in noisy and reverberant environments. Also, the subvector selection method that selects one phase difference unwrapping for given peaks in the initial DoA histogram may also be erroneous in the adverse environments. In this paper, we propose to modify the DoA histogram so that each frequency contributes to the DoA histogram equally by probabilistic voting to the DoA, which is equivalent to modify the histogram as

$$\boldsymbol{h}_{prop} = \sum_{k=1}^{K} \frac{1}{2A_k + 1} \hat{\boldsymbol{h}}_k, \quad (14)$$

where $\boldsymbol{h}_{prop}$ is the modified histogram. It can be considered to the midway between the traditional methods that disregards IPDs for the frequencies affected by spatial aliasing and the phase replication method that allows high frequencies provide multiple DoA candidates. It can also be interpreted as reducing the effect of the high frequency components where the speech signal may be weak.

Figure 2 shows a comparison between $\hat{\boldsymbol{h}} = \sum_{k=1}^{K} \hat{\boldsymbol{h}}_k$ with spectral masks only and $\boldsymbol{h}_{prop}$ with both spectral masks and probabilistic voting when the two sound source sources are located at -30 and 30 degrees, respectively. We can see that the true peaks are emphasized and the false peaks due to the spatial aliasing and interference are suppressed through the probabilistic voting of the DoA in the high frequencies.

We estimated the number and the DoAs of multiple sound sources using the peak search algorithm [15, 16] from the DoA histogram $\boldsymbol{h}_{prop}$. To smooth the DoA histogram and emphasize the peaks, we apply linear predictive coding (LPC) coefficient approach [16] to $\boldsymbol{h}_{prop}$ to produce $\tilde{\boldsymbol{h}}_{prop}$. In the smoothed histogram, the DoA corresponding to the highest peak $\hat{\theta}_1$ is regarded as the DoA for one source. If the next peak $\tilde{\boldsymbol{h}}_{prop}(\hat{\theta}_{i+1})$ is higher than a threshold, $\beta\tilde{\boldsymbol{h}}_{prop}(\hat{\theta}_i)$ for a factor β, the number of estimated sources is increased by 1 and $\hat{\theta}_{i+1}$ is regarded as the corresponding DoA. The peak search continues until the next peak is lower than the threshold or the maximum number of sources P_N is reached.

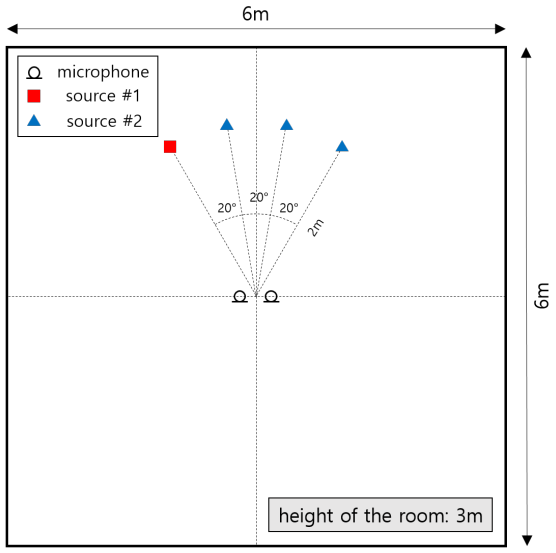


Figure 3: *The room configurations and the locations of the microphones and sound sources in the experiments.*

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 $6 \times 6 \times 3$ 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°, 2 m away from the center of two microphones. The second sound source was also 2 m away from the center of two microphones, and the DoAs were -10°, 10°, and 30°. 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 noises were 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 source signals with the SNRs from 0 dB to 20 dB at 5 dB intervals. The parameters were set as $K = 257$, $\lambda_{TH} = 5$ 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 $minRMSE$ [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 $RT60 = 200\ ms$.*

source#2		-10°			10°			30°			Average		
method		[13]	mask	Prop.	[13]	mask	Prop.	[13]	mask	Prop.	[13]	mask	Prop.
SNR	0 dB	21.51	15.23	**5.19**	11.71	7.59	**7.42**	5.99	9.74	**5.60**	13.07	10.85	**6.07**
	5 dB	22.04	12.35	**1.87**	9.78	7.41	**6.52**	5.61	5.84	**4.04**	12.47	8.53	**4.14**
	10 dB	17.60	9.11	**1.85**	7.71	6.22	**5.11**	5.46	4.02	**3.07**	10.26	6.45	**3.34**
	15 dB	11.29	6.67	**2.03**	7.00	6.08	**4.83**	5.32	**3.00**	3.21	7.87	5.25	**3.36**
	20 dB	8.87	5.23	**2.11**	6.79	5.55	**4.68**	5.26	**2.38**	3.38	6.97	4.39	**3.39**
Average		16.26	9.72	**2.61**	8.60	6.57	**5.71**	5.53	5.00	**3.86**	10.13	7.10	**4.06**

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 $SNR = 20\ dB$.*

source#2		-10°			10°			30°			Average		
method		[13]	mask	Prop.	[13]	mask	Prop.	[13]	mask	Prop.	[13]	mask	Prop.
$RT60$	1.0 s	22.63	15.60	**13.96**	10.52	8.91	**8.79**	**10.05**	13.44	10.66	14.40	12.65	**11.14**
	0.8 s	22.86	14.35	**12.00**	10.76	**8.42**	8.67	9.99	12.59	**9.92**	14.53	11.79	**10.19**
	0.6 s	22.98	11.64	**10.49**	10.56	8.34	**8.10**	9.85	11.38	**7.87**	14.46	10.45	**8.82**
	0.4 s	23.99	**7.82**	8.30	9.48	8.17	**7.33**	8.75	8.05	**5.48**	14.07	8.01	**7.04**
	0.2 s	12.14	5.23	**2.11**	6.35	5.55	**4.68**	5.49	**2.38**	3.38	7.99	4.39	**3.39**
Average		20.92	10.93	**9.37**	9.53	7.88	**7.51**	8.83	9.57	**7.46**	13.09	9.46	**8.12**

is two and the number of estimated sources is also two for the most of the cases, the $minRMSE$ is defined as the minimum of the $RMSE$s for the possible source assignments:

$$minRMSE = \min(RMSE_{12}, RMSE_{21}), \quad (15)$$

where

$$RMSE_{12} = \sqrt{\left((\theta_1 - \hat{\theta}_1)^2 + (\theta_2 - \hat{\theta}_2)^2\right)/2}, \quad (16)$$

$$RMSE_{21} = \sqrt{\left((\theta_2 - \hat{\theta}_1)^2 + (\theta_1 - \hat{\theta}_2)^2\right)/2}, \quad (17)$$

in which θ_1 and θ_2 denote the true DoAs of the two sources. If the number of estimated sources was 1, which was 0.7% of the cases, we used $\hat{\theta}_1$ instead of $\hat{\theta}_2$. If the number of estimated sources was more than 2, which was 1.4% of the cases, $minRMSE$ was computed for first two DoA estimates.

Table 1 shows the average $minRMSE$ depending on the SNR and the DoA of the second source when the $RT60$ was fixed at 200 ms. 1500 utterances were used in this experiment in total (10 utterances for source 1 $\times$ 10 utterances for source 2 $\times$ 5 SNRs $\times$ 3 positions of source 2). The 95% confidence intervals for the average of all the cases in the first experiment were 0.370, 0.448, and 0.266 for [13], that with spectral masks only, and the proposed method, respectively. We can see that the proposed method outperformed the phase replication method in [13] in all the cases, especially when two sources are located close to each other. It is noted that the performance of two source localization gets better when the DoA difference between two sources gets bigger, and when the sources are located near the broadside direction. These two effects made the $minRMSE$ not a monotonic function of the DoA for the second source. By applying spectral masks, we could obtain better performance than [13] except a few cases with low SNRs. The proposed method with both spectral masks and modified histogram with equal contribution from all frequency bins further improved the performance. Table 2 shows the performance according to the $RT60$ and the DoA of the second source when the SNR was fixed at 20 dB. As in the first experiment, 1500 utterances were used in total. The 95% confidence intervals for the average of all the cases in the second experiment were 0.363, 0.433, and 0.385, respectively. We can see that the proposed method showed superior performance for most of the cases. As in the first experiment, applying spectral masks improved performance for most of the cases and the probabilistic voting further enhance the performance. From the experimental results, we could confirm that the proposed spectral masks and the probabilistic voting enhanced the performance of the SSL using the phase replication method.

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)).

7. References

[1] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 6, pp. 1079–1093, 2016.

[2] J. Dmochowski, J. Benesty, and S. Affès, "On spatial aliasing in microphone arrays," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1383–1395, 2008.

[3] M. S. Amin, K. I. Ahmed, Z. R. Chowdhury *et al.*, "Estimation of direction of arrival (doa) using real-time array signal processing," in *2008 International Conference on Electrical and Computer Engineering*. IEEE, 2008, pp. 422–427.

[4] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 33–36.

[5] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a laplacian mixture model," *Digital Signal Processing*, vol. 21, no. 1, pp. 66–76, 2011.

[6] C. Kim and K. K. Chin, "Sound source separation algorithm using phase difference and angle distribution modeling near the target," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] C. Kim, A. Menon, M. Bacchiani, and R. Stern, "Sound source separation using phase difference and reliable mask selection selection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5559–5563.

[8] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, 2019.

[9] K. Itoh, "Analysis of the phase unwrapping algorithm," *Applied optics*, vol. 21, no. 14, pp. 2470–2470, 1982.

[10] R. Krämer and O. Loffeld, "Presentation of an improved phase unwrapping algorithm based on kalman filters combined with local slope estimation," in *ERS SAR Interferometry*, vol. 406, 1997, p. 253.

[11] V. V. Reddy and A. W. Khong, "Direction-of-arrival estimation of speech sources under aliasing conditions," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1–5.

[12] K. Chen, J. T. Geiger, and W. Kellermann, "Robust audio localization with phase unwrapping," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 471–475.

[13] K. Chen, J. Geiger, W. Jin, M. Taghizadeh, and W. Kellermann, "Robust phase replication method for spatial aliasing problem in multiple sound sources localization," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 195–199.

[14] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.

[15] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, 2008.

[16] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[18] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[19] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[20] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.