INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Weighted Spatial Covariance Matrix Estimation for MUSIC based TDOA Estimation of Speech Source

Chenglin Xu[1,2], *Xiong Xiao*[2], *Sining Sun*[3], *Wei Rao*[2], *Eng Siong Chng*[1,2], *Haizhou Li*[2,4]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3] School of Computer Science, Northwestern Polytechnical University, China
[4] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{xuchenglin,xiaoxiong}@ntu.edu.sg snsun@nwpu-aslp.org
{raowei,aseschng}@ntu.edu.sg haizhou.li@nus.edu.sg

Abstract

We study the estimation of time difference of arrival (TDOA) under noisy and reverberant conditions. Conventional TDOA estimation methods such as MUltiple SIgnal Classification (MUSIC) are not robust to noise and reverberation due to the distortion in the spatial covariance matrix (SCM). To address this issue, this paper proposes a robust SCM estimation method, called weighted SCM (WSCM). In the WSCM estimation, each time-frequency (TF) bin of the input signal is weighted by a TF mask which is 0 for non-speech TF bins and 1 for speech TF bins in ideal case. In practice, the TF mask takes values between 0 and 1 that are predicted by a long short term memory (LSTM) network trained from a large amount of simulated noisy and reverberant data. The use of mask weights significantly reduces the contribution of low SNR TF bins to the SCM estimation, hence improves the robustness of MUSIC. Experimental results on both simulated and real data show that we have significantly improved the robustness of MUSIC by using the weighted SCM.

Index Terms: Time Difference of Arrival (TDOA), Weighted Spatial Covariance Matrix (WSCM), MUltiple SIgnal Classification (MUSIC)

1. Introduction

Time difference of arrival (TDOA) estimation of acoustic sources is essential for a wide range of applications such as source localization and tracking [1, 2, 3, 4], teleconferencing systems [5], far-field speech recognition [6, 7, 8]. However, the performance of TDOA estimation degrades significantly in very noisy and reverberant environments. The robust TDOA estimation still needs to be studied in such challenging conditions.

Over the last few decades, the generalized cross correlation with phase transform (GCC-PHAT) method was widely used for TDOA estimation [9]. The limitation was that errors were accentuated where the signal power was low. Another popular approach was MUltiple SIgnal Classification (MUSIC) [10], which divided the spatial covariance matrix (SCM) to source subspace and noise subspace by the eigenvalue decomposition and then estimated the TDOA by searching the steering vectors orthogonal to the noise subspace. However, the SCMs of time-frequency (TF) bins dominated by noise and reverberation could alter the estimation of the TDOA in low SNR and high reverberant scenarios. To solve this problem, several approaches were proposed by selecting reliable TF bins dominated by directed sound. For example, Guo et al. [4] extracted reliable regions by tracking the envelopes of speech, reverberation and background noise. But the assumption of only one speech source in de-aliasing process limited the usability in real applications. In [11], a combination of noise-floor tracking, onset detection and coherence test was proposed to robustly identify time-frequency (TF) bins where only one source was dominant. However, some noisy TF bins might be wrongly selected as a result of fixed threshold, especially in very low SNR scenarios.

Inspired by neural network based masking methods for beamforming [12, 13, 14], we propose the weighted spatial covariance matrix (WSCM) based MUSIC method, named as WSCM-MUSIC, for robust TDOA estimation by selecting speech dominated TF bins through a long short term memory (LSTM) [15] based mask predictor. Specially, the mask predictor is trained using a large amount of simulated noisy and reverberant data to cover as many conditions as possible. The predicted mask is valued from 0 to 1 and indicates the possibility of each TF bin dominated by speech. Then the WSCM is computed for every TF bin enhanced by the estimated mask. In this way, the contributions of the noise and reverberation in the WSCM are heavily attenuated. The MUSIC is applied on the WSCM of each TF bin to obtain a pseudo spectrum. Finally, the TDOA is estimated by finding a peak from the summed pseudo spectrum over all TF bins to overcome the spatial aliasing ambiguity occurring at high frequencies. The experimental results on simulated and real conditions show that the performance of the proposed WSCM-MUSIC method is better than the GCC-PHAT [9], the learning based system [16], MUSIC [10] and its variant [11] in low SNR and high reverberant environments.

In Section 2, the signal model and the TDOA estimation problem are formulated. Our proposed WSCM-MUSIC method is introduced in Section 3. Section 4 shows the data simulation, experimental setup and results. We conclude the study in Section 5.

2. Problem Formulation

Considering a planar and circular array with M microphones in a 2D geometry, their observations in frequency domain are denoted by $X_m(n,f), m = 1, 2, \ldots, M$, in noisy and reverberant environment. The signal at microphone m is modeled as:

$$X_m(n,f) = \alpha_m S(n,f) e^{-j2\pi f \tau_m} + H_m(n,f) + V_m(n,f) \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2017-199

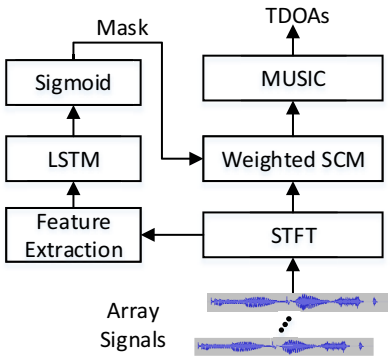

Figure 1: *A TDOA estimation architecture using WSCM.*

where n and f are the indexes of time frame and frequency bin. α_m is the attenuation factor due to propagation energy decay and channel effects. τ_m is the propagation time from the source location to the microphone m. And the TDOAs are $\tau_m - \tau_1, m = 2, \ldots, M$. $S(n, f)$ represents the sound source and $H_m(n, f)$ and $V_m(n, f)$ denote the reverberation and additive noise at microphone m. $V_m(n, f)$ may be correlated when the noise is directional, e.g., from an air conditioner or a computer.

As stated in [4], the phase of $|X_m(n, f)|$ was determined by $2\pi f\tau_m$ only if $\alpha_m S(n, f) \gg |H_m(n, f)| + |V_m(n, f)|$. It means that the estimated TDOA is close to its true value only if the TF bin is dominated by the direct sound. Thus only reliable parts dominated by speech should be extracted for TDOA estimation. In the low SNR and high reverberant case, the selection of reliable region is still a very challenging problem. Furthermore, due to the wide space of microphones, the spatial aliasing ambiguity may occur at high frequencies [17, 18]. Therefore, the robust TDOA estimation still needs further investigation.

3. WSCM-MUSIC Approach

The proposed WSCM-MUSIC approach, as shown in Fig. 1, can be summarized as follows. First, the reliable TF bins that carry the TDOA information are extracted. This is realized by implementing a LSTM based mask predictor to estimate the speech mask. Then the WSCM is computed for the eigenvalue decomposition in MUSIC algorithm. Furthermore, the summation is applied to the pseudo spectrum of every TF bin obtained from MUSIC to overcome the spatial aliasing problem at high frequencies. Finally, the TDOAs are estimated from the summed pseudo spectrum by finding peaks.

3.1. Mask Predicting

To estimate the WSCM, we firstly estimate a speech mask for each TF bin,

$$w(n, f) = f_{W_1}(X) \tag{2}$$

where W_1 represents the neural network parameters.

The structure of the LSTM based mask predictor is shown in Fig. 1. In this mask predictor, the log power spectrum and its dynamics (delta and acceleration) are used as input features. The utterance based cepstral mean normalization is applied on the log power spectrum features. The hidden activations of the LSTM layer are mapped to the speech mask through a sigmoid layer. In the sigmoid layer, the sigmoid function is applied on a linear affine transform to ensure the predicted masks between 0 to 1. Unlike previous mask prediction work with pooling operation over multiple channels [14], we train the mask predictor only using first channel speech data and estimate masks for all channels using the same predictor. In the training stage, since we have the separated clean and noise signals, the ideal binary masks are obtained to guide the training of the network through the cost function of mean square error.

3.2. Definition of WSCM

For the multi-channel signals shown in equation (1), the SCM of a TF bin is defined as

$$R_x(n, f) = E[X(n, f)X^H(n, f)] = R_s + R_h + R_v \tag{3}$$

where $E[\cdot]$ is the expectation operation, and ^{H} denotes conjugate transpose. $X(n, f) = [X_1(n, f), \ldots, X_M(n, f)]$ is the signals of all M channels in frequency domain. R_s, R_h and R_v are corresponding to the speech, reverberation and noise covariance matrixs.

Since R_s contains much spatial information and can not be directly obtained, it's a good way to estimation $\hat{R}_s$, which is close to the true R_s. Given the speech mask $\hat{w}(n, f) = \sqrt{w(n, f)/\sum_n w(n, f)}$, the WSCM can be computed as

$$\hat{R}_s(n, f) = E[\hat{w}(n, f)X(n, f)X^H(n, f)\hat{w}^H(n, f)] \tag{4}$$

3.3. TDOA estimation

Given the WSCM ($\hat{R}_s(n, f)$), we directly do eigenvalue decomposition. Since only one source signal is considered in this paper [1], the eigenvector (u_1) corresponding to the largest eigenvalue is obtained to span the signal subspace as $U_s(f) = [u_1]$ at frequency f. And the eigenvectors ($[u_2, \ldots, u_M]$) corresponding to other $M-1$ eigenvalues are orthogonal to the signal subspace and span the noise subspace as $U_n(f) = [u_2, \ldots, u_M]$. Since the steering vector ($a(\theta)$) from the true arrival direction also belongs to the signal subspace, the MUSIC algorithm works by constructing an arrival angle dependent power expression named as pseudo spectrum, and then searching for all steering vectors that are orthogonal to the noise subspace. The pseudo spectrum is defined as,

$$P(f, \theta) = \frac{1}{a^H(\theta)U_n(f)U_n^H(f)a(\theta)} \tag{5}$$

where θ is the arrival angle and $a(\theta)$ is the corresponding steering vector.

For speech signal, we always use several TF bins, e.g., 256 TF bins for speech sampling rate at $16kHz$. For each TF bin, we obtain a pseudo spectrum. The pseudo spectrum of all TF bins are summed together to overcome the spatial aliasing ambiguity occurring at high frequencies. Then the peaks are found from the summed pseudo spectrum, where the steering vectors of the arrival angles are orthogonal to the noise subspace.

$$P(\theta) = \sum_f P(f, \theta) \tag{6}$$

In practice, the steering vector of the arrival angle is never exactly orthogonal to the noise subspace, because there is noise and the obtained covariance is estimated on samples. The highest peak is selected as the desired direction of arrival (DOA) and the time delays are calculated from the estimated DOA with the array geometry and the planar wave assumption.

[1] If D source signals (less than microphone number) are considered, the eigenvectors corresponding to the largest D eigenvalues are obtained to compose the signal subspace.

Table 1*: The simulation settings for generating training and test data. All rooms are of 3m room heights.The "Distance" means distance between the array and the source.*

Training Data Simulation	
Speech	7861 utterances from WSJCAM0 training set
Room Size (m)	small (7×5), medium (12×10), large (17×15)
Distance (m)	near (1), far (2, 4, 6.5 for small, medium, large)
T60 (s)	0.1s to 1.0s with 0.1s step
SNR (dB)	uniformly sampled from 0dB to 20dB
Test Data Simulation	
Speech	538 utterances from WSJCAM0 et1 test set
Room Size (m)	small (6×4), medium (10×8), large (14×12)
Distance (m)	near (1), far (1.5, 3, 5 for small, medium, large)
T60 (s)	$0.3s, 0.6s, 0.9s$ for small, medium, large room
SNR (dB)	-10dB, 0dB, 10dB, 20dB
Real Test Data Recording	
Speech	64 utterances
Room Size (m)	small (6×4), large (10×7)
Distance (m)	1.5m and 3m for small, 6m for large room

4. Experiment and Analysis

4.1. Experimental Setup

4.1.1. Data Simulation and Recording

We conduct the experiments of TDOA estimation on simulated and real data. The data is simulated and recorded by a 8-channel circular array with a diameter of 20cm. To create the simulation data, the clean speech signals are convolved with the room impulse responses (RIRs) generated using the image method [19, 20] based on the given circular array geometry. Then the noises from Reverb Challenge 2014 [21] are added. The simulated data varies from different room sizes, source to array distances, reflection rates (resulting in different T60s), and SNR levels.

In the training data simulation, the $7,861$ clean utterances from the WSJCAM0 [22] training set are firstly convolved with various RIRs generated with different settings (as shown in Table 1). Before generating the RIR for an utterance, the room size and source to array distance are randomly decided from the given scenarios in Table 1. The T60 is also randomly selected from 0.1s to 1.0s with a 0.1s step. We assume that the source keeps static in the whole utterance. Then the randomly selected noises are added to each utterance by the SNR levels randomly chosen from 0dB to 20dB. We run the above simulation for 6 times resulting in $47,166$ training utterances from 360 DOA angles from 0 to 359 degree.

Similar to the training data generation, the 538 clean utterances from WSJCAM0 test set are convolved with the designed 18 scenarios (3 room sizes $\times$ 2 distances $\times$ 3 SNRs) [2]. For each scenario, the 360 DOA angles from 0 to 359 degree are randomly assigned to 360 test utterances. In the real test data recording, 64 utterances are spoken at 8 different DOA angles from 0 to 359 with a 45 degree step. We record these utterances in 2 room size and 3 source to array distance individually, as stated in Table 1.

4.1.2. Parameter Settings

Same as in [16], 588 dimensional GCC features in the learning based method are extracted with a window size of 0.2s and shift of 0.1s, which consist of $C_2^8 = 28$ pairs $\times$ 21 correlation coefficients [3]. We repeat the GCC based classification method for DOA estimation with the same setting in [16], named as GCC-CLASS in this paper. After obtaining the DOA, we convert it to TDOA based on the array geometry as comparison. To mitigate the mismatch problem, the same strategies of weighting GCC, HEQ [23] and max normalization in [16] are still applied in real test data scenario.

In the proposed WSCM-MUSIC method, the mask predicting network is configured with 771 input features, a LSTM based hidden layer with $1,024$ nodes and 257 output nodes. The STFT length is 512 samples and the frame window size and shift is 25ms and 10ms. The hamming window is applied. In addition, we conduct a TF bin selection scheme using noise-floor tracking, onset detection and coherence test in [11] as comparison. After selecting TF bins, the MUSIC method is applied on the SCM computed from the selected TF bins, named as SSCM-MUSIC in this paper. Furthermore, the classical GCC-PHAT [9] and MUSIC [10] methods are also compared.

Table 2: *The TDOA estimation results on the simulated data. The root mean square error (RMSE) in samples is used as evaluation metric. The RMSEs with different SNR, T60 and distance are reported.*

SNR	Method	RMSE (Samples)					
		T60=0.3s		T60=0.6s		T60=0.9s	
		1m	1.5m	1m	3m	1m	5m
-10dB	GCC-PHAT	1.44	2.21	1.35	3.32	0.90	3.26
	GCC-CLASS	3.37	3.71	3.71	4.49	3.34	4.95
	MUSIC	2.26	1.95	1.65	3.46	1.3	2.86
	SSCM-MUSIC	0.56	0.20	0.16	**0.22**	0.06	**0.18**
	WSCM-MUSIC	**0.22**	**0.14**	**0.03**	0.64	**0.03**	0.49
0dB	GCC-PHAT	0.57	0.69	0.56	0.88	0.56	0.85
	GCC-CLASS	1.41	1.27	1.37	2.86	1.09	3.71
	MUSIC	0.05	0.08	0.03	0.15	0.03	0.22
	SSCM-MUSIC	0.06	0.10	0.03	0.09	**0.02**	**0.08**
	WSCM-MUSIC	**0.05**	**0.07**	**0.03**	**0.08**	0.03	0.11
10dB	GCC-PHAT	0.56	0.45	0.56	0.50	0.56	0.50
	GCC-CLASS	0.14	0.11	0.24	0.86	0.26	1.89
	MUSIC	0.06	0.06	0.03	0.08	0.03	0.11
	SSCM-MUSIC	0.04	**0.05**	**0.02**	**0.05**	0.02	**0.05**
	WSCM-MUSIC	**0.04**	0.06	0.03	0.08	**0.02**	0.11
20dB	GCC-PHAT	0.56	0.39	0.56	0.49	0.56	0.50
	GCC-CLASS	0.10	0.14	0.05	0.25	0.03	0.79
	MUSIC	0.04	0.06	0.03	0.08	0.02	0.11
	SSCM-MUSIC	0.04	**0.04**	**0.02**	**0.04**	0.02	**0.04**
	WSCM-MUSIC	**0.04**	0.06	0.03	0.08	**0.02**	0.10

4.2. Results and Comparison on Simulated Data

We firstly evaluate our proposed WSCM-MUSIC approach in several simulated scenarios and compare it with other methods. The comparative results are shown in Table 2. We observe that the performances of GCC-PHAT, GCC-CLASS and MUSIC methods are serious affected by high level noise such as low SNR of -10dB. This verified our claim that the classic methods are not robust in challenging environment. In the GCC-CLASS method, the low SNR of -10dB scenario is not simulated in training stage, so the performance is not so good because of mismatch problem. In the SSCM-MUSIC method, the performance has been improved a bit by selecting reliable TF bins through noise floor tracking, onset detection and coherence test. It works well in moderate and high SNR conditions. But the performance is not as good as our proposed WSCM-MUSIC method in most low SNR conditions, because the bin selection in low SNR may be not reliable and some noisy TF bins may be selected.

The performance of the proposed WSCM-MUSIC method is better than others in most conditions, especially in low SNR conditions. Comparing with the MUSIC method, the main con-

[2] We only specify one T60 for a given room size.

[3] The array diameter is 0.2m, the maximum delay in samples is $0.2/340 * 16000 \approx 21$, so the centered 21 correlation coefficients are selected.

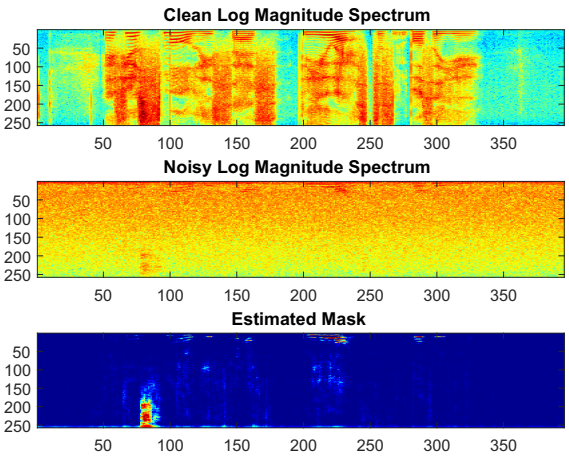

Figure 2: *The spectrum of clean signal and observed noisy signal, the estimated speech mask. This utterance ('c30c020k' from wsjcam0) is simulated in small room (6x4x3) with T60=0.3, SNR=-10dB, distance=5m and doa angle=15.*

tribution comes from the improved mask estimation. An example of the predicted mask and the summed pseudo spectrum in very challenging environment is shown in Fig. 2 and Fig. 3. The proposed WSCM-MUSIC estimates the arrival angle exactly same as true angle (=15), since the highest peak is at there. Although the MUSIC method has a peak at 15, the arrival angle is wrongly estimated as around 140 by finding highest peak in the MUSIC method. The wrong estimation is caused by the noise and reverberation in the estimation of the SCM. Our proposed WSCM-MUSIC solves the problem by filtering the noise and reverberation using predicted mask.

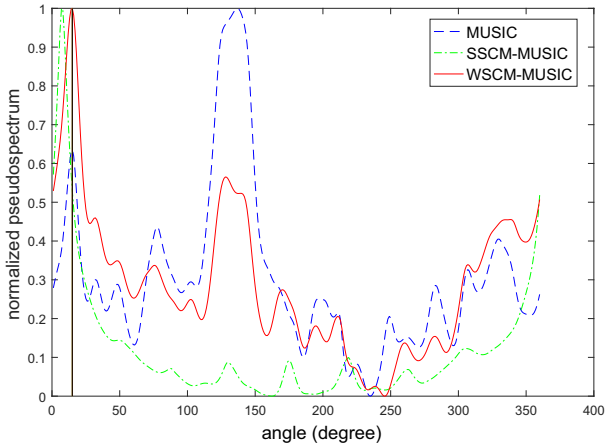

Figure 3: *The pseudo spectrum of MUSIC methods for utterance in Fig. 2. True doa angle is 15 marked with vertical line.*

4.3. Results and Comparison on Real Data

To evaluate the performance in real environment, we record some utterances using the setting as shown in Table 1. The results are summarized in Fig. 4. We observe that the MUSIC, SSCM-MUSIC and WSCM-MUSIC methods have almost same performance. The main reason is that there is little noise in the recordings and the spatial covariance estimation is not severely affected by noise. In order to evaluate the robustness of the proposed WSCM-MUSIC method in real environment, we add noise to the real recordings to the SNR level of 0dB. Fig. 5 shows that the performances of the SSCM-MUSIC and WSCM-MUSIC are better than MUSIC. The main contribution comes from the reliable TF bin selection that is essential in noisy and reverberant environment. Furthermore, the proposed WSCM-MUSIC is better than the SSCM-MUSIC. It means that the TF bin selection is more reliable using the LSTM based mask predictor than the scheme in [11]. The main reason is that the mask predictor leverages the learning ability of LSTM from a large amount of data covered many challenging conditions.

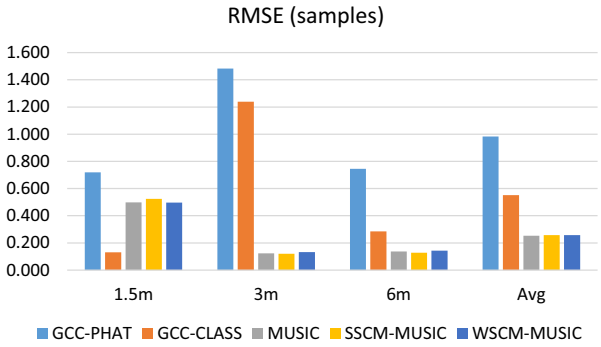

Figure 4: *The TDOA estimation on real data. The distances in a small room are 1.5m and 3m. The distance of 6m is used in a large room recording.*

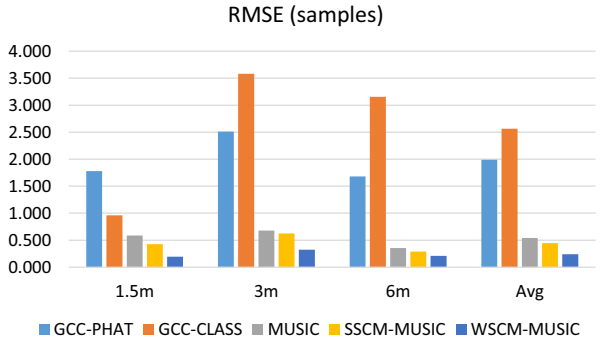

Figure 5: *The TDOA estimation on real data by adding noise to SNR of 0dB, since the recordings don't have too much noise.*

5. Conclusions

In this paper, the WSCM-MUSIC method is proposed for the robust TDOA estimation in very challenging conditions. The problem and solution were presented for the robust TDOA estimation. Several GCC and SCM based systems were also evaluated and compared. Experimental results showed that the proposed WSCM-MUSIC method worked well in very noisy and heavily reverberant environment. The advantage and robustness of the LSTM based mask predictor were verified comparing to the previous proposed TF bin selection method of combining noise-floor track, onset detection and coherence test. Since the LSTM based mask predictor is trained for only one speaker without leveraging TDOA information, the mask predictor for more than one speaker will be considered in the future work. And the TDOA information will be investigated in the training process of the mask predictor.

6. Acknowledgement

Thanks to Dr. Zhao Shengkui for the TF bin selection code. This work is supported by the DSO funded project MAISON DSOCL14045, Singapore.

7. References

[1] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[2] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[3] X. Zhong and J. R. Hopgood, "A time–frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2356–2370, 2015.

[4] Y. Guo, X. Wang, C. Wu, Q. Fu, N. Ma, and G. Brown, "A robust dual-microphone speech source localization algorithm for reverberant environments," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, 2016.

[5] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3d sound localization system with miniature microphone array for virtual reality," in *Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on*. IEEE, 2012, pp. 1853–1857.

[6] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5745–5749.

[7] X. Xiao, S. Watanabe, E. S. Chng, and H. Li, "Beamforming networks using spatial covariance features for far-field speech recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–6.

[8] X. Xiao, C. Xu, Z. Zhang, S. Zhao, S. Sining, S. Watanabe, L. Wang, L. Xie, D. L. Jones, E. S. Chng, and H. Li, "A study of learning based beamforming methods for speech recognition," *The 4th International Workshop on Speech Processing in Everyday Environments (CHIME 2016)*, pp. 26–31, 2016.

[9] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[10] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[11] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2287–2291.

[12] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 196–200.

[13] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016.

[14] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3246–3250.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[17] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[18] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[19] 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.

[20] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.

[21] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[22] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 81–84.

[23] A. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.