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

# A Phase-Based Time-Frequency masking for multi-channel speech enhancement in domestic environments

*Alessio Brutti*[1], *Antigoni Tsiami*[2], *Athanasios Katsamanis*[3], *Petros Maragos*[2,3]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]National Technical University of Athens, Athens, Greece
[3]Athena Research and Innovation Center, Maroussi, Greece
brutti@fbk.eu, antsiami@cs.ntua.gr, nkatsam@cs.ntua.gr

## Abstract

This paper introduces a novel time-frequency masking approach for speech enhancement, based on the consistency of the phase of the cross-spectrum observed at multiple microphones. The proposed approach is derived from solutions commonly adopted in spatial source separation and can be used as a post-filter in traditional multi-channel speech enhancement schemes. Since it is not based on a modeling of the coherence of diffuse noise, the proposed method complements traditional post-filters implementations, targeting non diffuse/coherent sources. It is particularly effective in domestic scenarios where microphones in a given room capture interfering coherent sources active in adjacent rooms.

An experimental analysis on the DIRHA-GRID corpus shows that the proposed method considerably improves the signal-to-interference-ratio and can be used on top of state-of-the-art multi-channel speech enhancement methods.

**Index Terms**: speech enhancement, microphone array, post-filter.

## 1. Introduction

Thanks to the recent advances in Automatic Speech Recognition (ASR), a variety of voice-enabled applications are emerging today, including solutions for home-automation and domestic scenarios in general. With respect to more traditional office-related scenarios, the domestic context presents additional challenges: the high variability of the acoustic conditions and the presence of competitive speech sources or interferers in other rooms of the home. This source of disturbance is not commonly targeted in literature: it cannot be addressed like diffuse background noise neither it represents an actual source separation problem. As an example, Voice Activity Detection (VAD) methods relying on statistical models of the speech signals may be significantly affected by these competitive noise components, unless very articulated algorithms are implemented [1].

Multi-channel speech enhancement is typically achieved using a Minimum Variance Distortionless Response (MVDR) beamformer followed by a single channel Wiener (post)-filter. This combination has been shown to be equivalent to the optimal multi-channel Minimum Mean Square Error (MMSE) enhancement [2]. A variety of different formulations for the post-filter are available in literature [3, 4, 5] where the main open issue is a proper estimation of the Power Spectral Density (PSD) of the residual noise and of target, in order to derive the Wiener filter. The state of the art solutions typically rely on a model of the diffuse sound coherence to establish which Time-Frequency (T-F) bins of the spectrogram of the recorded signals are dominated by the target speech and which are noisy. These solutions are not totally suitable for the domestic application scenario where noise includes speech sources being active in other rooms or other non-diffuse sounds.

We propose to attack this particular problem by applying a multi-channel T-F masking based on the cross-spectrum phase information. Following strategies already used in source localization in presence of multiple sources [6], we introduce a multi-channel formulation of the MMSE post-filter by defining a probabilistic soft association between the target source and each T-F point, based on the distribution of the cross-spectrum phase. From a more general perspective, the proposed approach can be seen as an alternative way to solve the problem of evaluating local SNR in each time-frequency bin to build the ideal Wiener post-filter, offering at the same time the opportunity to extend the approach to source separation. Finally, since it does not employ any assumption on the noise properties and on the structure of the noise field coherence, it is particularly suitable to address coherent interfering sources, complementing this way the traditional post-filter schemes.

The phase of the cross-spectrum (or interaural time difference) has been used for source separation using two microphones in DUET [7] and in [8] in combination with source-model strategies. Similarly, in [9, 10] a cross-spectrum phase based post-filter for a two-microphone beamformer is presented. Analogous concepts are applied to microphone arrays in [11], where the phase-based T-F mask is defined for sub-bands rather than in a bin-wise manner. A slightly different approach is investigated in [12], where the variance of the estimated direction of arrival is used in the post-filter estimation.

The paper is organized as follows. Section 2 introduces the multi-channel speech enhancement problem while Section 3 presents the proposed approach. The experimental analysis is detailed in Section 4. Finally, Section 5 concludes the paper with future works and final remarks.

## 2. Problem formulation

In the targeted multi-room scenario, while a source emits a speech signal $s(t)$ in a given room, in presence of background noise and other noise sources, an interferer $i_r(t)$ is probably active in an adjacent room $r$. Assuming that $M$ microphones monitor the target enclosure, at each microphone $m$, $m = 1, \ldots, M$, the received signal can be modeled as:

$$x_m(t) = h_m * s(t) + \sum_{r=1}^{R} h_{m,r} * i_r(t) + \eta_m(t) \qquad (1)$$



http://dx.doi.org/10.21437/Interspeech.2016-150

where: $h_m$ is the RIR between the source and the microphone $m$, $h_{m,r}$ is the RIR between the interferer in room $r$, $r = 1, \ldots, R$, and the microphone $m$ and $\eta(t)$ is the environmental background noise. The term $\sum_r h_{m,r} * i_r(t)$ models the presence of competitive speech sources in other rooms. Without loss of generality we assume that a single interferer is active in each room (possibly as a result of the linear combination between the multiple interfering sources).

### 2.1. Multi-channel speech enhancement

Let us approximate all interfering components as a single uncorrelated noise field:

$$\nu_m(t) = \sum_{r=1}^{R} h_{m,r} * i_r(t) + \eta_m(t). \quad (2)$$

This approximation is acceptable in many cases as the lack of line-of-sight in combination with the multi-path propagation is expected to highly de-correlate the sound emitted by sources outside the room. Therefore, Eq. 1 can be simplified as:

$$x_m(t) = h_m * s(t) + \nu_m(t). \quad (3)$$

Applying the Short-Time Fourier Transform (STFT) and assuming that impulse responses are time invariant, Eq. 3 can be formulated in the frequency domain as follows:

$$\mathbf{X}(t,k) = \mathbf{H}(k)S(k) + \mathbf{V}(t,k), \quad (4)$$

where:

$$\mathbf{X}(t,k) = [X_1(t,k), \ldots, X_M(t,k)]^T, \quad (5)$$
$$\mathbf{H}(k) = [H_1(k), \ldots, H_M(k)]^T, \quad (6)$$
$$\mathbf{V}(t,k) = [V_1(t,k), \ldots, V_M(t,k)]^T, \quad (7)$$

and $X_m(t,k)$, $H_m$ and $V_m(t,k)$ are the STFT of $x_m(t)$, $h_m$ and $\nu_m(t)$ respectively. In anechoic conditions, $H_m(k)$ reduces to a simple delay and attenuation [13]:

$$H_m(k) = D_m(k) = \alpha_m e^{-jw_k\tau_m}, \quad (8)$$

where $\tau_m$ is the Time Difference of Arrival (TDOA) between a generic channel $m$ and the reference channel $m = 1$, while $w_k$ is the angular frequency in radiants of the $k$-th bin. $\mathbf{D}(k) = [D_1(k), \ldots, D_M(k)]^T$ is the array steering vector [2]. It has been shown that the optimum multi-channel MMSE speech enhancement filter can be decomposed as [14, 2]:

$$\mathbf{W}_{\text{opt}} = \left[\frac{\phi_{ss}}{\phi_{ss} + \phi_{\nu\nu}}\right] \frac{\mathbf{\Phi}_{\nu\nu}^{-1}\mathbf{D}}{\mathbf{D}^H\mathbf{\Phi}_{\nu\nu}^{-1}\mathbf{D}}, \quad (9)$$

where the right-end component is an MVDR beamformer [15]. The component:

$$\mu(t,k) = \left[\frac{\phi_{ss}}{\phi_{ss} + \phi_{\nu\nu}}\right] \quad (10)$$

in Eq. 9 is the single-channel Wiener filter. $\phi_{ss}$ and $\phi_{nn}$ are the PSD of the target signal and of the noise after the beamforming. The filter is often estimated assuming gaussianity of the noise or imposing models of the noise field coherence. Finally, the estimation of the target signal is obtained as:

$$\mathbf{Y}(t,k) = \mathbf{W}_{\text{opt}}^H(t,k)\mathbf{X}(t,k) \quad (11)$$

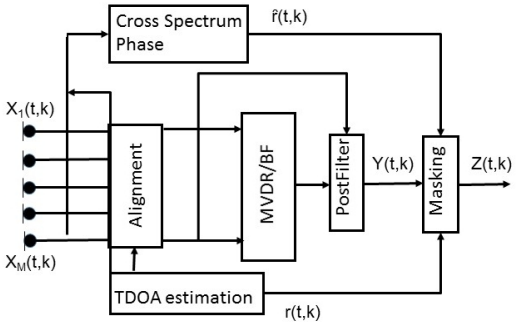



Figure 1: Block diagram of the proposed T-F masking scheme.

## 3. Proposed T-F Masking

In our target scenario, the assumption introduced above does not always hold. In particular, the term $\sum_r h_{m,r} * i_r(t)$, although rather uncorrelated, will never match the characteristics of typical noise fields, in particular when only one interferer is active. As a consequence, state-of-the-art solutions fail to completely remove these interference sources. Therefore, we introduce a novel definition of the Wiener filter based on the multi-dimensional distance between the expected phase of the cross-spectrum and the observed one.

The phase of the cross-spectrum between channel $m$ and the reference channel 1 at frequency bin $k$ and time instant $t$ is defined as [16]:

$$e^{-j\psi_m(t,k)} = \frac{X_m(t,k)X_1^*(t,k)}{|X_m(t,k)||X_1(t,k)|}. \quad (12)$$

Similarly to Eq. 8, in anechoic and noise-free condition, Eq. 12 simplifies as a linear phase:

$$e^{-j\psi_m(t,k)} = D_m(k) = e^{-jw_k\tau_m}, \quad (13)$$

In the same direction as in [10], we define the multi-channel phase error as [6]:

$$\epsilon(t,k) = \|\hat{\mathbf{r}}(t,k) - \mathbf{r}(t,k)\|_2 \quad (14)$$

where $\hat{\mathbf{r}}(t,k)$ is the estimated multi-channel phase:

$$\hat{\mathbf{r}}(t,k) = [e^{-j\psi_1(t,k)} \cdots e^{-j\psi_M(t,k)}] \quad (15)$$

and $\mathbf{r}(t,k)$ is the ideal linear phase given the target source spatial position:

$$\mathbf{r}(t,k) = [e^{-j2\pi f_k\tau_1} \cdots e^{-j2\pi f_k\tau_M}]. \quad (16)$$

Following the same principle as in [10], it makes sense to assume that those T-F bins where $\epsilon(t,k)$ is small correspond to bins with high SNR and low reverberation. The T-F masking is defined as the probability that a given T-F bin belongs to the target source given the TDOA vector $\boldsymbol{\tau} = [\tau_1, \ldots \tau_M]$:

$$\hat{\mu}(t,k) = p\left(\hat{\mathbf{r}}(t,k)|\boldsymbol{\tau}\right). \quad (17)$$

Assuming a multivariate gaussian distribution of the observation vectors, with frequency dependent variance, the mask can be computed as:

$$\hat{\mu}(t,k) = e^{-\epsilon(t,k)^2/\sigma_k^2} \quad (18)$$

where $\sigma_k$ accounts for the frequency dependent phase variance and has been empirically defined as:

$$\sigma_k = \frac{\beta}{(0.5 + f_k)}. \quad (19)$$

Basically $\sigma_k$ counterbalances the implicit low-pass effect of the phase error. The parameter $\beta$ controls the trade-off between distortion and interference removal. Small values of $\beta$ result in a very aggressive masking, high values of the parameter limit the action of the soft-masking (ideally for $\beta = \infty$ there is no masking at all).

The resulting T-F mask can be applied as post-filter after the beamformer or on top of a traditional post-filters to further improve the quality of the signal (as done in [17] for source separation):

$$\mathbf{Z}(t,k) = \hat{\mu}(t,k)\mathbf{Y}(t,k) \quad (20)$$

# 4. Experimental Analysis

We evaluated the proposed method on the DIRHA-GRID corpus [18], a multi-microphone and multi-room simulated database. The corpus contains a set of 225 acoustic scenes of 1-minute duration observed by 40 microphone distributed in a real apartment. Fig. 2 shows the layout used in the experiments with microphone positions. The star-like arrays in the Kitchen and Livingroom (top left and bottom left rooms in Fig. 2) are mounted on the ceiling. Each acoustic scene includes short English commands from the GRID database [19], and typical non-speech home noises. Each acoustic event occurs randomly in time and in space in any of the microphone-equipped rooms. In this dataset, a single speech events occurs in the apartment at each time, with other non-speech events (e.g., typical domestic noises, radio, etc.) possibly overlapping in time. Background noise consist of real noise recorded in the apartment (e.g., sound coming inside the apartment from an open window).

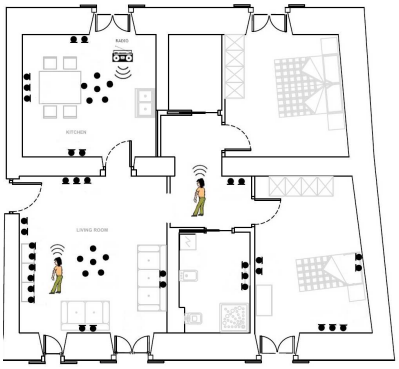


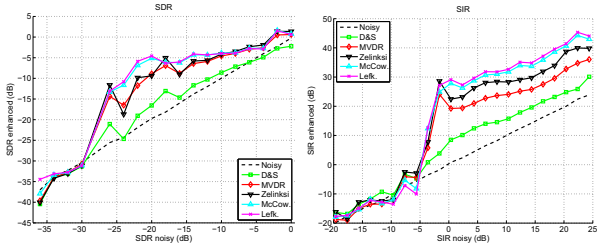
Figure 2: Layout of the apartment used in the experimental analysis.

While events may occur everywhere, the evaluation is limited to the Kitchen and the Livingroom, where the circular ceiling arrays allow implementing multi-channel enhancement algorithms. For each target room, the goal is to enhance speech events occurring inside, attenuating as much as possible noise and events occurring in other rooms. In practice, for each room the target signal is the noise-free sequence of close-talking GRID commands occurring inside the room. Signals are sampled at 16kHz and processed in 64ms windows with 16ms step.

We consider a set of metrics directly computed on the enhanced signals and adopted in the CHIME evaluation campaigns [20, 21] for blind source separation: Signal to Distortion Ratio (SDR) Signal to Interference Ratio (SIR). The metrics are computed using the "bss_eval 3.0" tool. The proposed method is compared and combined with state-of-the-art multi-channel speech enhancement techniques: Delay-and-Sum (D&S) beamforming, MVDR, Zelinksi [3], McCowan [4], Lefkimmiatis [5].

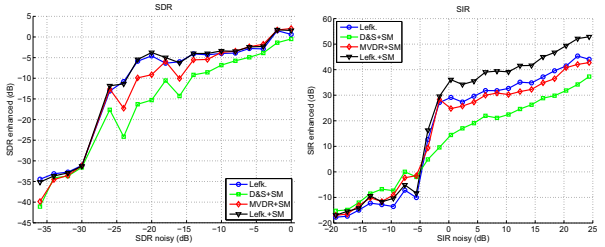
## 4.1. Results with oracle TDOA

To better understand the potential of the proposed method, in this first analysis we assume that oracle TDOA are available when the target source is active. Fig. 3 reports the speech enhancement performance in terms of SDR and SIR of the state-of-the-art algorithms as a function of the SDR and SIR of the noisy signal. The post-filters of [4] and [5] are particularly effective in terms of SDR in the central range. All approaches are effective in removing the interferers as soon as the SIR is above 0dB and the target source is dominant, with again [4] and [5] the best performing. Note that the SDR is particularly low because the target signal is the close-talking (noise-free and reverberation-free) signal.




(a) SDR (b) SIR

Figure 3: Improvement in terms of SDR and SIR provided by the state-of-the-art algorithms.




(a) SDR (b) SIR

Figure 4: Improvement in terms of SDR and SIR provided by the proposed algorithm with $\beta = 1.5$. SM refers to the proposed masking approach.

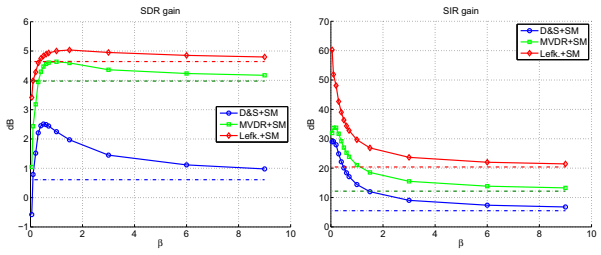
Fig. 4 compares the performance of the proposed T-F masking when it is used as post-filter with D&S and MVDR and when it is employed as a second post-filter in combination with [5]. $\beta = 1.5$ in this experiment. It is interesting to observe that the proposed approach is particularly effective in terms of SIR, providing an improvement of approximately 10 dB over the speech enhancement scheme it is combined with. Conversely, for low SIR cases there seems to be no benefit at all. For what concerns SDR, there is basically no improvement

Table 1: Improvement in dB on the full DIRHA-GRID dataset of the addressed speech enhancement algorithms. $\beta = 1.5$ in the masking. SM refers to the proposed masking approach.

| TDOA | metric | D&S | MVDR | Zelinski | McCowan | Lefk. | D&S+SM | MVDR + SM | Lefk.+SM |
|---|---|---|---|---|---|---|---|---|---|
| Oracle | SDR | 0.61 | 3.97 | 4.43 | 4.66 | 4.64 | 1.88 | 4.48 | 4.98 |
| | SIR | 5.50 | 12.17 | 16.27 | 19.48 | 20.39 | 11.90 | 18.41 | 26.81 |
| Estim. | SDR | -0.06 | 3.86 | 4.31 | 4.27 | 4.35 | 0.26 | 4.14 | 4.72 |
| | SIR | 1.92 | 11.61 | 15.20 | 17.48 | 18.15 | 5.05 | 14.53 | 23.91 |

when the proposed masking is combined with [4] and [5] and only a marginal gain is observed when it is directly used as post-filter after the MVDR and/or D&S. Table 1 summarizes the improvement over the noisy signals as average on the full dataset for $\beta = 1.5$.

To conclude, Fig. 5 shows the performance as a function of the parameter $\beta$ which, controlling the phase standard deviation in Eq. 19, determines the aggressiveness of the masking. The figure shows that, as expected, high SIR gains can be obtained only at the cost of a reduced SDR. Note that for high values of $\beta$ the effect of the masking is null and the performance tends to what obtained with the baseline methods. In terms of SDR the best performance is achieved for $\beta = 1.5$. Note that in this analysis we consider the optimum $\beta$ on average over all conditions. Ideally, $\beta$ should adapt, depending on: the amount of noise and reverberation in the signals, the number of channels and the accuracy of the TDOA estimation.




(a) SDR (b) SIR

Figure 5: Improvement in terms of SDR and SIR provided by the proposed masking as a function of the parameter $\beta$.

### 4.2. Results with estimated TDOA

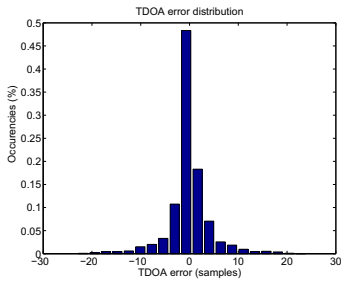
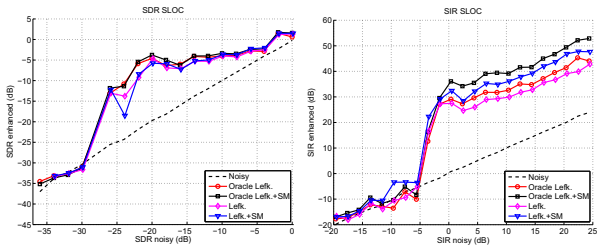



Figure 6: Distribution of the TDOA estimation error.

We now analyze the performance of the proposed approach when the source spatial position, and hence the TDOA, is estimated using the localization algorithm described in [22], which features an outlier elimination method that identifies and removes microphones providing erroneous information. On the DIRHA-GRID corpus, the proposed localization algorithm delivers 48.6% of estimates within an error of less than 50 cm per utterance, confirming that the localization task is particularly challenging. For what concerns the corresponding TDOA estimates, which is the information actually used in the speech enhancement algorithm, Fig. 6 shows the error distribution in samples: note that 77% of the estimation errors are within $\pm 1$ sample.

Fig. 7 shows the enhancement performance of the proposed method in combination with Lefkimmiatis, compared with performance obtained when oracle TDOA are available. Although a degradation is observed, the proposed enhancement scheme still brings benefits in terms of both metrics and is robust against noisy TDOA estimates. The bottom part of Table 1 reports the average performance on the full data set, considering also other solutions and combinations of methods.




(a) SDR (b) SIR

Figure 7: Enhancement performance when TDOA are automatically estimated. $\beta = 1.5$ in the T-F masking.

## 5. Conclusions

This paper presented a novel approach for the definition of a soft T-F masking (or post-filter) for multi-channel speech enhancement. Initially developed to tackle the effects of speech events occurring in adjacent rooms in domestic contexts, the proposed method results effective also in removing non-diffuse interfering sources and background noise.

Future work will address the introduction of an adaptive phase variance in Eq. 19, which, properly modeling the phase probability distribution based on the operational conditions, would optimize the filter behavior. In this direction, further improvements could be obtained by removing the gaussianity assumption and employing more articulated modeling of the phase distribution. For instance, following recent trends in multi-channel enhancement, neural networks could be used [23, 24, 25]. A further limitation of the proposed approach is the anechoic phase modeling in Eq. 13. Adopting an echoic modeling, possibly based on some awareness of the acoustic propagation in the monitored enclosure, could lead to better performance in presence of mild to high reverberation.

# 6. References


[1] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proc. EUSIPCO*, 2015.

[2] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[3] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 2578–2581 vol.5.

[4] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov 2003.

[5] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7, pp. 657–666, 2007.

[6] A. Brutti and F. Nesta, "Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs," *Computer Speech and Languages*, vol. 27, pp. 660–682, 2013.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[8] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," in *Proc. INTERSPEECH*, 2008, pp. 419–422.

[9] G. Shi and P. Aarabi, "Robust digit recognition using phase-dependent time-frequency masking," in *International Conference on Multimedia and Expo, ICME*, vol. 3, July 2003, pp. III–629–32 vol.3.

[10] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 4, pp. 1763–1773, Aug 2004.

[11] P. Pertilä, "Online blind speech separation using multiple acoustic speaker tracking and time-frequency masking," *Computer Speech and Language*, vol. 27, no. 3, pp. 683 – 702, 2013.

[12] I. Tashev and A. Acero, "Microphone array post-processor using instantaneous direction of arrival," in *Proceedings of International Workshop on Acoustic, Echo and Noise Control IWAENC*, September 2006.

[13] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, Dec. 2003.

[14] K. U. Simmer, J. Bitzer, and C. Marro, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, ch. Post-Filtering Techniques.

[15] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct 1987.

[16] C. H. Knapp and G. C. 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.

[17] R. M. Toroghi, F. Faubel, and D. Klakow, "Multichannel speech separation with soft time-frequency masking," in *SAPA-SCALE*, 2012.

[18] M. Matassoni, R. F. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Proc. INTERSPEECH*, 2014, pp. 1613–1617.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of Acoustic Society of America*, Nov. 2006.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[21] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, Aug. 2012.

[22] A. Tsiami, A. Katsamanis, P. Maragos, and G. Potamianos, "Experiments in acoustic source localization using sparse arrays in adverse indoors environments," in *Proc. EUSIPCO*, 2014.

[23] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[24] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Integration of speech enhancement and recognition using long-short term memory recurrent neural network," in *Proc. Interspeech*, 2015.

[25] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *12th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., 2015.