Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition

Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr

Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition

Sunit Sivasankaran, Emmanuel Vincent
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France
{sunit.sivasankaran,evincent}@inria.fr

Dominique Fohr
Université de Lorraine, CNRS, LORIA
F-54000 Nancy, France
dfohr@loria.fr

***Abstract*—We investigate the effect of speaker localization on the performance of speech recognition systems in a multispeaker, multichannel environment. Given the speaker location information, speech separation is performed in three stages. In the first stage, a simple delay-and-sum (DS) beamformer is used to enhance the signal impinging from the speaker location which is then used to estimate a time-frequency mask corresponding to the localized speaker using a neural network. This mask is used to compute the second order statistics and to derive an adaptive beamformer in the third stage. We generated a multichannel, multispeaker, reverberated, noisy dataset inspired from the well studied WSJ0-2mix and study the performance of the proposed pipeline in terms of the word error rate (WER). An average WER of 29.4% was achieved using the ground truth localization information and 42.4% using the localization information estimated via GCC-PHAT. Though higher signal-to-interference ratio (SIR) between the speakers was found to positively impact the speech separation performance, equivalent performances were obtained for mixtures with lower SIR values when the speakers are well separated in space.**

***Index Terms*—Multichannel speech separation, WSJ0-2mix reverberated**

I. Introduction

Speech captured by a distant microphone is corrupted by reverberation and noise. In a typical home scenario, it is often further distorted by interfering speakers. This problem, referred to as the speech separation problem or the cocktail party problem, has been studied for more than 20 years [1], [2]. With the advent of neural networks, it has regained the attention of the community.

In presence of multiple speakers different time-frequency bins are dominated by different speakers and the goal is to estimate a time-frequency mask for each speaker. The problem has been addressed in both single-channel and multichannel contexts. Single-channel approaches include clustering-based methods such as deep clustering [3] and deep attractor networks [4] where a neural network is trained to cluster together the time-frequency bins dominated by the same speaker. In another approach, the speakers are estimated iteratively [5] using neural networks with permutation-invariant training criteria [6].

In multichannel scenarios, the usual approach is to estimate the second-order statistics (covariance matrices) of all speech and noise sources and to derive a beamformer to separate the speakers [7], [8]. The separation quality will therefore depend on the covariance matrix estimates. Different methods have been proposed to estimate the target speech and noise covariance matrices. In [9], the phase differences between the microphones encoding speaker location information are exploited as input features to train a deep clustering based neural network for speech separation. Explicit speaker location estimates have also been employed. In [10] and [11], the speaker is first localized and the microphone signal is beamformed towards the speaker. The beamformed signal is used to estimate a mask corresponding to the localized speaker which is then used to estimate the covariance matrices. A similar approach is proposed in [12] where the so-called speech presence probability (SPP) is estimated using speaker location information with a minimum Bayes risk detector. The speech and noise statistics are then derived from the SPP.

Speech separation algorithms are often evaluated using speech enhancement metrics such as the signal-to-distortion ratio (SDR) and the perceptual estimation of speech quality (PESQ) metric [9], [12] and, in limited cases, using automatic speech recognition (ASR) metrics [11], [13], [14]. A related study on analyzing the impact of localization errors on WER was done in [14], but under limited acoustic conditions and vocabulary size.

In this paper, we provide the following contributions. We create a new multichannel, multispeaker, reverberated, noisy dataset which extends the original WSJ0-2mix single-channel, non-reverberated, noiseless dataset [3] to the strong reverberation and noise conditions and the Kinect-like microphone array geometry used in CHiME-5 [15]. This allows us to use the real noise captured as part of the CHiME-5 dataset, thereby making

This work was made with the support of the French National Research Agency, in the framework of the project VOCADOM "Robust voice command adapted to the user and to the context for AAL" (ANR-16-CE33-0006). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

the simulated dataset quite realistic and challenging. On this dataset, we perform speech separation using either the ground truth location of the speakers or the location estimated by the established generalized cross-correlation phase transform (GCC-PHAT) algorithm [16], and we evaluate the resulting ASR performance on the separated speech.

The rest of the paper is organized as follows. Section II introduces the proposed framework for speech separation using speaker localization information. Section III explains the procedure used to simulate the dataset. Section IV describes the experimental procedure and the obtained results are discussed in Section V. We conclude in Section VI.

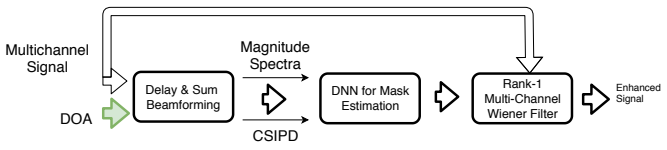
II. Speech separation using localization information

A. Signal model

The multichannel signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ captured at I microphones can be expressed as $\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t)$, where $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{Ij}(t)]^T$ is the spatial image of source j, i.e., the signal emitted by the source and captured at the microphones. Similar to [8], the microphone index and the time index are denoted by i and t, respectively, and J is the total number of sources. This general formulation is valid for both point sources as well as diffuse noise. For point sources such as human speakers, the spatial image can be expressed as a linear convolution of the room impulse response (RIR) $\mathbf{a}_j(t,\tau) = [a_{1j}(t,\tau), \ldots, a_{Ij}(t,\tau)]^T$ and a single-channel source signal $s_j(t)$: $\mathbf{c}_j(t) = \sum_{\tau=0}^{\infty} \mathbf{a}_j(t,\tau) s_j(t-\tau)$. Under the narrowband approximation, $\mathbf{c}_j$ in the time-frequency domain can be written as $\mathbf{c}_j(t,f) = \mathbf{a}_j(t,f) s_j(t,f)$.

Our objective is to estimate the spatial image of each source given its (known or estimated) spatial location. An overview of our speaker location guided speech separation system is shown in Fig. 1. This system comprises three steps: delay-and-sum (DS) beamforming, mask estimation, and adaptive beamforming. We detail each of these steps in the three subsections below.

Fig. 1: Speech separation pipeline using rank-1 MWF as the adaptive beamformer.



B. DS beamforming

Given the spatial location of source j in far-field, the corresponding time difference of arrival between a pair of microphones i and i' can be obtained as:

$$\text{TDOA}(i, i', j) = \frac{d_{ii'} \cos(\theta_{ii'j})}{c} \tag{1}$$

where $\theta_{ii'j}$ is the direction of arrival (DOA) of the source with respect to the microphone pair (i, i'), $d_{ii'}$ is the distance between the two microphones, and c is the velocity of sound.

A steering vector with respect to a reference microphone (in the following, microphone 1) can be computed as $\tilde{\mathbf{d}}_j(f) = [1, e^{-2j\pi(q_{ij}-q_{1j})v_f/c}, \ldots, e^{-2j\pi(q_{Ij}-q_{1j})v_f/c}]^T$ where ν_f is the continuous frequency (in Hz) corresponding to the frequency bin index f. The output of a simple DS beamformer for source j can then be obtained as

$$\hat{\mathbf{c}}_{j,\text{DS}}(t,f) = \tilde{\mathbf{d}}_j(f)^H \mathbf{x}(t,f). \tag{2}$$

where ^{H} denotes Hermitian transposition.

The localized speaker j is more prominent in $\hat{\mathbf{c}}_{j,\text{DS}}$ than in $\mathbf{x}$. We hence use $\hat{\mathbf{c}}_{j,DS}$ to compute a time-frequency mask corresponding to that speaker.

C. Time-frequency mask estimation

The magnitude spectrum of $\hat{\mathbf{c}}_{j,DS}$ and its phase difference with respect to the reference microphone are used as inputs to a neural network that estimates the time-frequency mask corresponding to the localized speaker.

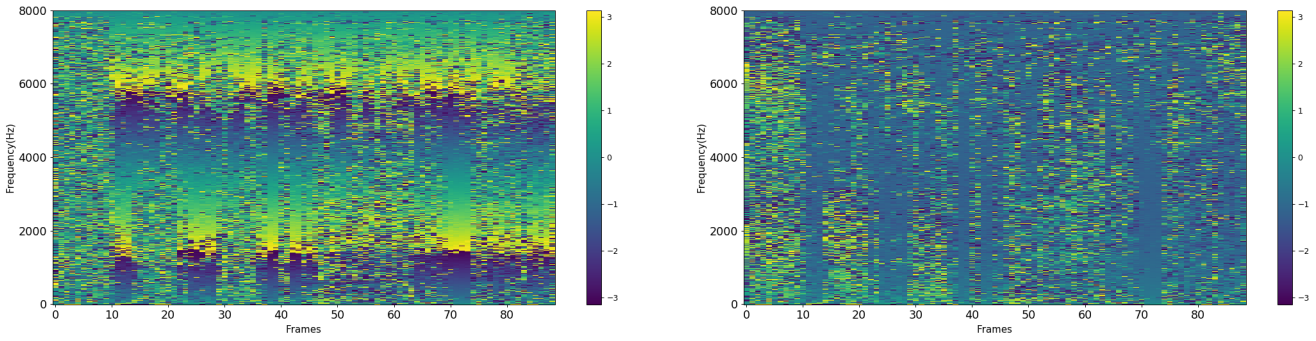
Using the phase difference between $\hat{\mathbf{c}}_{j,DS}$ and a reference microphone as a feature may not seem intuitive at first and requires further justification. Figure 2 shows the information captured by this phase difference. Fig. 2a shows the phase difference of the direct component (without reverberation) of a source between two microphones placed at a distance of 0.226 m in the presence of noise. The phase difference is perturbed in the time-frequency bins dominated by noise.

Fig. 2b shows the phase difference of the beamformed signal with respect to the signal at the reference microphone. The phase difference in the time-frequency bins dominated by speech is now zero, and a clear speech-like pattern can be observed in these bins. In the presence of reverberation, the speech patterns are less clearly visible before or after DS beamforming. Nevertheless, we argue that the phase difference contains useful information regarding the source which can be leveraged by a neural network in addition to the magnitude spectrum of the DS beamformer output in order to estimate a better time-frequency mask.

Since the phase difference is defined modulo 2π only, we use its cosine and sine as features, as used in [18] for speaker localization and in [9] for speech separation. We refer to these features as cosine-sine interchannel phase difference (CSIPD) features. These features are given as inputs along with the magnitude spectrum of $\hat{\mathbf{c}}_{j,DS}$ to train a neural network to estimate a mask. We highlight the fact that the dimension of the input features to train the mask estimation network does not depend on the number of microphones. In theory, we can use the same network for any number of microphones in the array.

D. Adaptive beamforming

The mask $M_j(t,f)$ output by the neural network for a given source j can be used to estimate the covariance matrix of that



(a) (b)

Fig. 2: Phase difference in presence of noise before DS beamforming (2a) and after DS beamforming (2b)

source as

$$\mathbf{\Sigma}_j(t,f) = \alpha\mathbf{\Sigma}_j(t-1,f) + (1-\alpha)M_j(t,f)\mathbf{x}(t,f)\mathbf{x}^H(t,f) \quad (3)$$

where α is a forgetting factor. Similarly, the noise covariance matrix $\mathbf{\Sigma}_n$, which includes the statistics corresponding to all other speakers and background noise, can be estimated as

$$\mathbf{\Sigma}_\mathbf{n}(t,f) = \alpha\mathbf{\Sigma}_\mathbf{n}(t-1,f) + (1-\alpha)(1-M_j(t,f))\mathbf{x}(t,f)\mathbf{x}^H(t,f). \quad (4)$$

An adaptive beamformer, i.e., a beamformer depending on the above statistics rather than the spatial location, is applied to the mixture signal $\mathbf{x}(t,f)$ to recover the sources. The output of the beamformer is $\mathbf{w}^H(t,f)\mathbf{x}(t,f)$. Different beamformers can be defined based on the chosen optimization criterion [8], [19]. In this work we consider the generalized eigenvalue (GEV) beamformer [20], the speech distortion weighted multichannel Wiener filter (SDW-MWF) [21], and the rank-1 constrained multichannel Wiener Filter (R1-MWF) [22].

III. Dataset

The data for this work is based on a multichannel, reverberated, noisy version of the WSJ0-2mix dataset [1]. The original WSJ0-2mix dataset introduced in [3] was created by mixing pairs of speakers from the WSJ0 corpus, and contains 20 k, 5 k, and 3 k training, development, and test mixtures, respectively. Each mixture contains two different speakers speaking for a variable duration. In this work, the "max" version of the dataset is used where the length of mixed signals is the maximum of the length of individual signals.

In our experiments we emulate the recording conditions of the CHiME-5 corpus which was recorded using Microsoft Kinect devices. For each pair of speech signals in WSJ0-2mix, we simulate room impulse responses (RIRs) using the RIR Simulator [23] for two distinct spatial locations with a minimum DOA difference of 5°. The room dimensions and the reverberation time (RT60) are randomly chosen in the range of $[3-9]$ m and $[0.3-1]$ s. The two speech signals are convolved with these RIRs and mixed at a random signal-to-interference ratio (SIRs) in the range of $[0-10]$ dB.

Real multichannel noise captured as part of the CHIME-5 dataset is then added with a random SNR in the range of $[0-10]$ dB. To obtain noise segments, the ground truth speech activity detection (SAD) labels from Track 3 of the DIHARD-II speaker diarization challenge [24] are used, as these are more reliable than the SAD labels originally provided in CHiME-5. The noise signals in the training, development, and test sets are taken from different CHiME-5 sessions. The noise is realistic and non-stationary in nature and makes the speech separation task very challenging. A reverberated dataset based on WSJ0-2mix was reported earlier in [25], but it does not contain any noise.

IV. Experimental settings

DNN to estimate the mask: Mask estimation is done in the time-frequency domain. The short time Fourier transform (STFT) of the 4-channel signal was computed using a sine window of length 100 ms and a shift of 50 ms resulting in a frequency dimension of 801. The input to the mask estimation network was of dimension 2403: it comprises the magnitude spectrum of the DS signal, as well as the cosine and sine of the phase differences as detailed in Section II-C, each of which is of dimension 801. A 2-layer Bi-LSTM network containing 801 hidden units was trained to estimate the mask corresponding to the reverberated component of the localized speaker. No dereverberation was performed. Adam was used as the optimizer.

ASR system: For each separation method tested, the ASR system was trained on the enhanced training set using accurate senone alignments obtained from the underlying clean single-speaker utterances. The acoustic model (AM) was a 15-layer time-delayed neural network (TDNN) trained using the lattice-free maximum mutual information criterion [26]. 40 dimensional Mel frequency cepstral coefficients along with 100-dimensional i-vectors were used as input features.

[1]The code to recreate the dataset can be found here: https://github.com/sunits/Reverberated_WSJ_2MIX

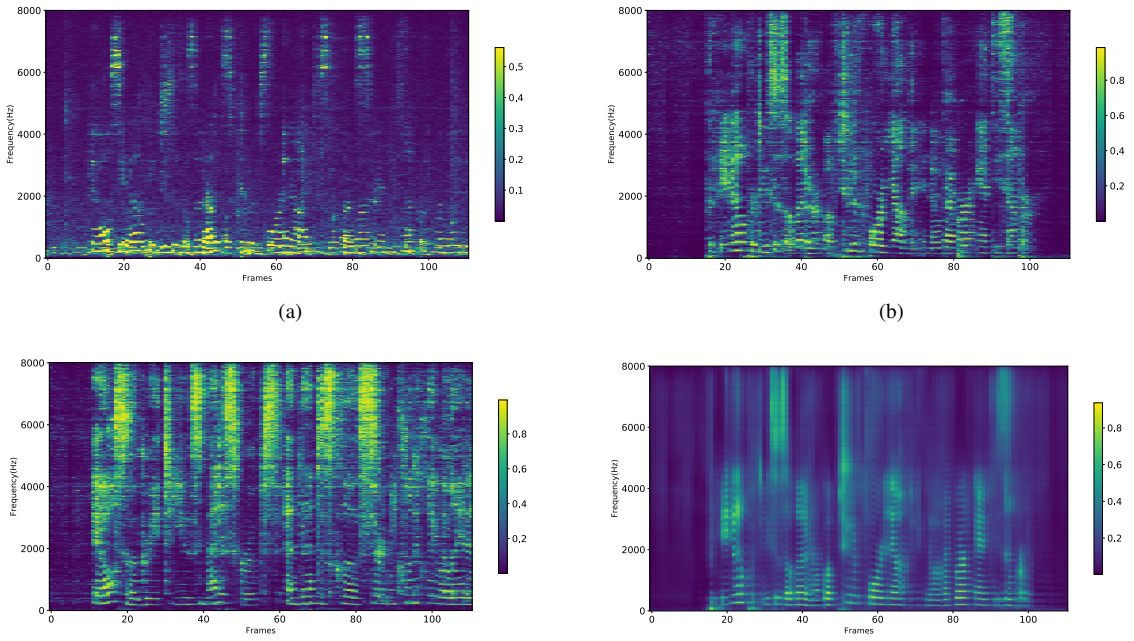


Fig. 3: (3a) Spectrogram of the 2 speaker mixture. (3b) True mask of the localized speaker. (3c) True mask of the non-localized speaker and (3d) The estimated mask of the localized speaker

Computing WER metrics for the mixture: In every mixture we perform ASR only for the speaker who spoke for the longest duration. This was done so that the insertion errors corresponding to the speaker who spoke for a shorter duration do not effect the ASR performance. In a typical ASR system, this is handled by a voice activity detector / endpointing system which is not the focus of this work.

Estimating location using GCC-PHAT: Experiments were conducted using both ground truth DOA values as well as using the DOA values estimated by GCC-PHAT. In the case of GCC-PHAT, peaks in the angular spectrum are assumed to correspond to the DOAs of the sources. The top two peaks are chosen and the peak which is closest to the true DOA is taken as the estimated DOA. Since GCC-PHAT works using 2 microphones, only the first and the last microphone of the array which are placed at a distance of 0.226 m are used.

V. Results

Figure 3a shows the spectrogram of a signal containing 2 speakers. The final mask estimated using the neural network (Fig 3d) can be observed to match that of the intended speaker in Fig 3b while ignoring the bins corresponding to the interfering speaker whose mask is shown in Fig 3c.

TABLE I: Baseline WER (%) achieved on single-speaker or two-speaker mixtures before enhancement/separation. All results reported in this paper are with reverberated speech.

1 **speaker**	1 **speaker + noise**	2 **speakers + noise**
12.2	23.6	58.2

Table I shows the baseline ASR performance before separation. It can be observed that background noise and overlapping speech severely degrade performance.

TABLE II: ASR performance after speech separation.

	GEV	**R1-MWF**	**SDW**
Using True DOA (%)	30.9	29.4	29.6
DOA with GCC-PHAT (%)	43.2	42.4	42.4

Table II shows the ASR results obtained on noisy two-speaker mixtures after speech separation. An average WER of 29.4% was obtained using the ground truth DOA, a relative improvement of 49% with respect to the system without source separation. This is close to the ASR performance for a single speaker with noise (23.6%) as shown in Table I, showing that DOA information can indeed help source separation. The performance dropped to 42.4% when the DOA was estimated using GCC-PHAT, indicating that erroneous DOA estimates decrease the separation quality[2]. In all our experiments R1-MWF beamformer outperformed the widely used GEV beamformer.

The impact of the DOA differences between the speakers and SIR on the ASR performance are shown in Table III. Better performances were consistently observed for signals with better SIR values and also for mixtures containing speakers who are well separated in space (Ex: $> 50^{\circ}$). It is interesting to note that better speech separation performances

[2]Speech separation model and AM trained using true DOA values were used since the corresponding models trained using estimated DOAs performed poorly.

TABLE III: Impact of SIR and DOA difference on ASR performance.

DOA diff vs SIR	< -5 **dB**	$[-5:0]$ **dB**	$[0:5]$ **dB**	> 5 **dB**
$< 10°$	67.0	43.2	25.7	26.3
$[10:30]°$	58.3	32.6	24.7	20.5
$[30:50]°$	60.0	32.0	23.4	22.2
$> 50°$	56.6	29.2	21.7	19.4

can be obtained for mixtures containing relatively lower SIR if the speakers are well separated in space compared to mixtures containing speakers who are spatially close to each other. For example, a WER of 21.7% was obtained for speech mixtures with SIR values between $[0:5]$ dB and DOA difference of $> 50°$, a relative improvement of 17.5% compared to mixtures with SIR > 5 dB and DOA difference $< 10°$.

VI. Conclusion

We conducted the first analysis of the impact of speaker localization accuracy on speech separation performance in challenging two-speaker, reverberated, noisy scenarios, as measured by the resulting ASR performance. To do so, we created a new dataset by reverberating WSJ0-2mix and mixing it with real CHiME-5 noise, and made the corresponding code publicly available. We found that the ASR performance depends more on the SIR of the speakers, with lower WERs for signals with higher SIR. The angular distance between the DOAs of the speakers was also found to have an impact, with better WERs for signals whose speakers exhibit a larger difference in DOAs.

References

[1] G.J. Brown and Martin Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, pp. 297–336, 1994.

[2] DeLiang Wang and Guy J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[3] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*. 2016, pp. 31–35, IEEE.

[4] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *ICASSP*. 2017, pp. 246–250, IEEE.

[5] Keisuke Kinoshita, Lukas Drude, Marc Delcroix, and Tomohiro Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *ICASSP*. Apr. 2018, pp. 5064–5068, IEEE.

[6] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[7] Michael Brandstein and Darren Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing. Springer-Verlag, Berlin Heidelberg, 2001.

[8] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[9] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, Feb. 2019.

[10] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *ICASSP*, Apr. 2018, pp. 36–40.

[11] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-Channel Overlapped Speech Recognition with Location Guided Speech Extraction Network," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 558–565.

[12] Maja Taseska and Emanuël A. P. Habets, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 60, Aug. 2017.

[13] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *ASRU*, Dec. 2017, pp. 437–444.

[14] Hendrik Barfuss and Walter Kellermann, "On the impact of localization errors on HRTF-based robust least-squares beamforming," in *DAGA*, Mar. 2016, pp. 1072–1075.

[15] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, Sept. 2018, pp. 1561–1565.

[16] 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, Aug. 1976.

[17] Tobias Menne, Ilya Sklyar, Ralf Schlüter, and Hermann Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," in *ICASSP*, May 2019.

[18] Sunit Sivasankaran, Emmanuel Vincent, and Dominique Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Interspeech*, Sept. 2018.

[19] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[20] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.

[21] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[22] Ziteng Wang, Emmanuel Vincent, Romain Serizel, and Yonghong Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, May 2018.

[23] Emanuël A. P. Habets, "RIR-Generator: Room impulse response generator," Feb. 2018.

[24] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Interspeech*, Graz, Austria, June 2019.

[25] Zhong-Qiu Wang, Jonathan Le Roux, and John R. Hershey, "Multi-Channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP*. Apr. 2018, pp. 1–5, IEEE.

[26] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI.," in *Interspeech*, 2016, pp. 2751–2755.