The Choice of Reference Channel in Channel Alignment and Channel Selection

Ingo Stengel[a], Karin Pietruska[a] and Matthias Wölfel[a]

[a] *Karlsruhe University of Applied Sciences, Moltkestraße 30, Karlsruhe, 76133, Germany*

Abstract

Channel alignment based on the generalized cross correlation with phase transform (GCC-PHAT) is part of many multichannel speech processing procedures including the channel selection procedure based on multichannel cross-correlation coefficients (MCCC). Despite the wide application of the GCC-PHAT approach for channel alignment, little has been reported on how the choice of reference channel might affect alignment accuracy and subsequent processing steps when microphones are coarsely distributed. The present research investigates alignment accuracy with random selection of a reference channel in relation to heuristic selection of a reference channel using the GCC-PHAT approach for time difference of arrival (TDOA) estimation and subsequent MCCC based channel selection. Results show that the procedure for reference channel selection effects both: the accuracy of channel alignment as well as results of the subsequent channel selection procedure. Findings suggest that the choice of reference channel should not be left to chance in distributed microphone arrays in order to optimize processing steps following channel alignment.

Keywords

Channel selection, multichannel speech processing, microphone array

1. Introduction

In various contexts the auditory information of a scene is recorded by several spatially distributed microphones, so called microphone arrays. In film production or sports broadcasting, spaced microphone arrays are used to create an immersive audio experience and to separate sound sources of interest from ambient noise [1, 2]. Likewise, in conference rooms or lecturing halls microphone arrays have proven useful to enhance sound signals that emanate from the current speaker while reducing noise from spatially distinct locations.

The estimation of time differences of arrival (TDOA) of sound signals at different microphones forms a critical first step in many techniques employed in microphone array processing for noise reduction [3], speaker localization [4, 5], channel selection [6] or speech enhancement [7]. First introduced more than half a century ago [8], the generalized cross correlation technique remains a widely applied method for TDOA estimation in near field and far-field scenarios [9]. The cross-correlation technique for TDOA estimation takes two signals as input and finds the time lag between the two signals that maximizes the value of the cross-correlation function. In the generalized cross correlation technique [8], an additional weighting function, also referred to as filtering, is applied to the cross-correlation. This paper focuses on the PHAT-weighting function, a filtering approach that has proven particularly useful for TDOA estimation in indoor settings that are characterized by signals with different forms of reverb [10]. Throughout the years, research has dedicated much attention on expanding and optimizing the GCC-PHAT approach. Only recently, a subband analysis with GCC-PHAT has yielded improved accuracy in TDOA estimates in relation to the classic approach [11]. To date, the GCC-PHAT is widely applied in multichannel signal processing and often constitutes one of the first steps when combining multiple signals. Channel alignment based on TDOA estimation with

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland
ingo.stengel@h-ka.de (I. Stengel); karin.pietruska@h-ka.de (K. Pietruska); matthias.woelfel@h-ka.de (M. Wölfel)

GCC-PHAT constitutes also the first step of the multichannel cross-correlation coefficient (MCCC) procedure for channel selection first introduced by Kumatani et al. (2011) [6].

Selecting a subset of channels of a microphone array for further multichannel processing remains of key interest in present research, particularly regarding voice-based assistance systems or conferencing systems that integrate signals received by a dispersed microphone array. As research has shown, adding more channels, particularly when they have a low signal to noise ratio, may not improve but instead substantially decrease the performance of an automatic speech recognition system [12]. Throughout the last decades, different approaches for channel selection have been introduced including classic signal to noise ratio estimation [13], class separability of phonemes [14], multichannel cross-correlation coefficients [6], cepstral distance [15] and neural network posterior probability models [16]. Notably, channel alignment remains an integral processing step in many of these multichannel approaches also including end-to-end ASR models, particularly with masked based neural beamforming [17].

We focus on the MCCC approach for channel selection, due to a suggested decrease in computational complexity compared to the use of an automatic speech recognizer (ASR) for channel selection [6, 14] and its use of the GCC-PHAT approach for channel alignment [6]. The MCCC approach aims at discarding low quality, noisy channels by building on the assumption that noise is uncorrelated to the speech signal of interest. Setting information on the spatial correlation among signals in relation to their variance, the MCCC algorithm implements a channel selection procedure that in combination with beamforming has yielded similar word error rates compared to a close distance microphone while focusing on computational efficiency. Channel alignment is implemented in the first processing step of the MCCC algorithm in order to optimize the accuracy and computational efficiency of the subsequently computed spatial correlations among channels [6]. Despite the wide use of the GCC-PHAT in the first step of channel alignment, little is known on how the choice of reference channel affects alignment accuracy and subsequent processing steps. This is of particular interest for microphone arrays that are spaced coarsely across the recording room with variable inter-microphone distances. This coarse setting differs profoundly from much of previous research that focused on linear, evenly spaced microphone arrays [18].

The present research aimed at investigating the effect of the choice of reference channel on TDOA accuracy for channel alignment based on the GCC-PHAT approach. Moreover, follow-up effects of the choice of reference channel for alignment on the overall results of the MCCC based channel selection procedure were examined. Random choice of reference channel was compared to a choice of reference channel based on a delay heuristic. Short-Time Objective Intelligibility (STOI) scores for each channel served as an independent speech intelligibility measure for the achieved channel rankings [19]. The effects were examined on data recorded in indoor settings with microphones distributed on a table or stand. The microphone locations were similar to a distribution that can be expected in ad-hoc microphone arrays when recording business meetings or seminars with the smartphones of meeting participants. The use of real data was critical in the approach as synthetic data are known to generate results that are often not replicable in realistic indoor settings.

2. Methods

2.1. Data

In order to investigate the present research questions, we aimed at using data recorded in indoor settings with unobstructed microphones spaced across a table or located on a stand. The VOiCES corpus provides recording conditions that fulfill these requirements along with spatial information that allowed the approximation of inter-microphone distances of a subset of the microphones used in the recordings [20]. We constrained the analysis to data recorded in room 3 (size: 7.6 by 7.6 m) with a foreground loudspeaker angle of 90 degrees. In the 90 degrees position, the loudspeaker is in line with the microphones of interest (azimuth angle = -90 degrees), causing a maximal time delay between subsequent microphones. None of the distractor noise loudspeakers were active. Prerecorded speech from LibriVox recordings was played by the foreground loudspeaker in room 3 equipped with basic furniture including a table, chairs, a shelf as well as a refrigerator. 20 microphones of different type (studio microphones, lavalier microphones, MEM microphones) were placed at different locations within the room. In the present paper, we constrained the microphones included to microphones that

were unobstructed, positioned either on a table or stand and located in front of the foreground speaker box. Inter-microphone distances were approximated by taking the difference scores of the indicated distance of each microphone to the foreground speaker box. Table 1 lists the included 7 microphones with information on the type of microphone, their location with respect to the foreground speaker box. Data were recorded with a PreSonus StudioLive RML32AI digital mixer and PreSonus Capture recording software and all channels were sampled synchronously with a sampling frequency of 16 kHz [20]. Table 2 and Table 3 depict the approximated inter-microphone distances as well as the expected differences in TDOA values in terms of samples given a sampling frequency of 16 kHz.

Table 1

Type, model and location of microphones included in the analysis. Microphone model and location descriptions are based on the documentation of the VOiCES corpus [20]

ID	Type	Model	Location
01	studio	SHURE SM58	close on table
02	lavalier	AKG 417L	close on table
03	studio	SHURE SM58	mid distance on table
04	lavalier	AKG 417L	mid distance table
05	studio	SHURE SM58	far distance on stand
06	lavalier	AKG 417L	far distance on stand
16	bar	ATR4697	mid distance on table

Table 2

Approximated inter-microphone distances in centimeters based on the given distance information of each microphone to the foreground speaker box. Height differences are not adequately represented in the calculated inter-microphone distances. Distance values in the original VOiCES corpus are indicated in inches without positions after the decimal point. Indicated inter-microphone distances are therefore broad approximations.

	ID 01	ID 02	ID 03	ID 04	ID 05	ID 06	ID 16
ID 01	0	0	201	201	544	544	104
ID 02	0	0	201	201	544	544	104
ID 03	201	201	0	0	343	343	97
ID 04	201	201	0	0	343	343	97
ID 05	544	544	343	343	0	0	439
ID 06	544	544	343	343	0	0	439
ID 16	104	104	97	97	439	439	0

Table 3

Expected inter-microphone delays in samples based on the approximated inter-microphone distances and a sampling frequency of 16 kHz. Positive difference values indicate that the respective channel in the column is delayed with respect to the reference microphone ID denoted by the row label. Conversely, negative values indicate that the channel denoted by the column label was located more closely to the sound and was therefore ahead in time compared to the channel denoted by the row label.

	ID 01	ID 02	ID 03	ID 04	ID 05	ID 06	ID 16
ID 01	0	0	94	94	253	253	49
ID 02	0	0	94	94	253	253	49
ID 03	-94	-94	0	0	160	160	-45
ID 04	-94	-94	0	0	160	160	-45
ID 05	-253	-253	-160	-160	0	0	-205
ID 06	-253	-253	-160	-160	0	0	-205
ID 16	-49	-49	45	45	205	205	0

2.2. Analysis 1: Reference Channel on TDOA

This first analysis investigated the effect of the choice of reference channel on the accuracy of channel alignment. Time differences of arrival (TDOA) for each channel with respect to a chosen reference channel were estimated with the generalized cross-correlation with PHAT weighting (GCC-PHAT). First introduced by Knapp and Karter (1976) [8], the generalized cross-correlation function denoted by $R_{km}(\tau)$ takes two microphone signals k, m as an input and computes the cross-correlation of the filtered versions of these two input signals. When applying the PHAT weighting, these filters consist of the phat weighting function denoted by $\psi_{km}(\omega)$ as described by the following equations:

$$R_{km}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_{km}(\omega) X_k(\omega) X_m(\omega)^* e^{j\omega\tau} d\omega \tag{1}$$

$$\psi_{km}(\omega) = \frac{1}{|X_k(\omega) X_m{}^*(\omega)|} \tag{2}$$

The maximum of $R_{km}(\tau)$ is the lag value estimated by the GCC-PHAT function that corresponds to the relative delay between the input signals k, m.

$$\widehat{\tau_{km}} = \arg\max \mathrm{R_{km}}\,(\tau) \tag{3}$$

In order to examine the effect of the choice of reference channel on alignment accuracy, random selection of one reference channel was compared to the selection of a reference channel based on a delay heuristic. Let k be the total number of channels. The delay heuristic attempts to estimate the channel located closest to the sound source by taking each channel k_r and computing the relative delays of all remaining k-1 channels with regard to channel k_r. The result is a delay matrix $\boldsymbol{D}$ of dimension k x k, whereby k is equal to the total number of channels. Each row of $\boldsymbol{D}$ contains the GCC-PHAT delay estimates $\widehat{\tau_{k_r k_l}}$ in relation to one specific reference channel k_r. The delay matrix $\boldsymbol{D}$ is a hollow matrix in which the diagonal values are all zeros as the relative delay of a channel k_r to itself is always zero.

$$D_{k,k} = \begin{pmatrix} \hat{\tau}_{1,1} & \hat{\tau}_{1,2} & \cdots & \hat{\tau}_{1,k} \\ \hat{\tau}_{2,1} & \hat{\tau}_{2,2} & \cdots & \hat{\tau}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\tau}_{k,1} & \hat{\tau}_{k,2} & \cdots & \hat{\tau}_{k,k} \end{pmatrix} \tag{4}$$

The first row of the delay matrix ***D*** contains the estimated time delays $\hat{\tau}$ of each channel with respect to channel 1. The second row of the delay matrix contains the delays of each channel with respect to channel 2 up until row *k* with reference channel *k*. The value of $\hat{\tau}$ is of positive sign when the respective channel is delayed to the reference channel of the respective row and negative if it is ahead of the reference channel of the respective row. More precisely, if the value of $\hat{\tau}_{1,2}$ is of positive sign, the signal of channel 2 was delayed with respect to channel 1. The heuristic aims at choosing the channel as the reference for the alignment procedure to which all other channels are delayed. Due to the possibility of maxima of the generalized cross correlation function $R_{km}(\tau)$ that might result from signal reflections or noise, the TDOA estimate may in some cases fail to reflect the ground truth time difference between two channels *k*,*m*. The heuristic therefore adopts the channel as reference channel for alignment with the maximal number of $\hat{\tau}$ values of positive sign within the respective delay matrix row. This means, based on the GCC-PHAT TDOA estimates, the maximal number of channels are delayed with respect to the reference channel.

Delay Heuristic:

Take each row $\boldsymbol{d}_{i*}$ of delay matrix ***D*** and compute the sum of the outputs of the sign function for each row entry.

$$g(d_{i*}) = \sum_{i=1}^{j} sgn(d_{ij}) \tag{5}$$

Take the index *i* of the row that maximizes the output of $g(d_{i*})$. This index *i* denotes the row of delay matrix ***D*** that contains the delays with respect to the reference channel k_{ref} selected by the delay heuristic for channel alignment:

$$k_{ref} = \underset{d_{i*} \in \mathcal{D}}{\operatorname{argmax}}\, g(d_{i*}) \tag{6}$$

2.3. Analysis 2: Reference Channel on MCCC Channel Selection

The second analysis investigates potential follow-up effects of the choice of reference channel for alignment on the results of the channel selection procedure based on the MCCC algorithm [6]. As described below, channel alignment based on TDOA estimates by GCC-PHAT constitutes the first step of the MCCC channel selection algorithm. Following the alignment procedure, the covariance matrix ***S*** is computed for each sample to capture the spatial correlations among channels.

$$S_{k,k} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_k^2 \end{pmatrix} \tag{7}$$

The MCCC score ρ_k for a specific sample is calculated by computing the determinant of the covariance matrix ***S*** and dividing it by the product of the diagonal elements s_i^2 of ***S***, which denote the variance of the signal within the respective channel i.

$$\rho_k = \frac{det[\mathbf{S}_k]}{\prod_{i=1}^{k} s_i^2} \quad (8)$$

For detailed mathematical notations, please refer to the paper by Kumatani et al. (2011) [6]. In the following, the channel selection algorithm based on MCCC values is briefly summarized:

1. TDOA estimation based on GCC-PHAT
2. Channel Alignment of the *k* signals
3. Denote all the channels in the search space as K_c
4. Find set K_s of K_c-1 signals with highest MCCC value
5. Remove k_i that was not included in set of K_c-1 with highest MCCC value from search space
6. Go to step (3.) if MCCC value of K_c is larger than the MCCC value of the subset K_s, $K_c > K_s$

The smallest set K_s of channels that can be retained by the algorithm consists of two channels. At least two channels are needed to compute a spatial correlation and thus the MCCC value. By saving the channels that are excluded in subsequent rounds of the algorithm, we receive a channel ranking from worst quality to the two best quality channels as ranked by the MCCC algorithm.

In the present analysis, for each of the n=128 utterances of the VOiCES corpus, a channel ranking based on the MCCC algorithm was computed and compared to the ranking of the channels based on the Short-Time Objective Intelligibility (STOI) scores [19]. The STOI score has a range from 0 to 1 with higher scores representing increased speech intelligibility. Table 4 displays the STOI Scores for the first sample utterance for each of the k=7 channels as well as their distance to the foreground speaker-box. The LibriSpeech source recordings of the respective utterance served as the non-degraded signal for the STOI-score computation.

Table 4

STOI scores for each channel (=microphone) of sample 1 along with distance of each microphone to the foreground speaker. The original LibriVox signal served as the undegraded reference for computation of STOI scores.

Mic ID	Distance to Speaker (cm)	STOI Score
01	170	0.59
02	170	0.49
03	371	0.37
04	371	0.31
05	714	0.26
06	714	0.23
16	274	0.39

3. Results

3.1. Analysis 1: Reference Channel on TDOA

The present analysis investigated the effect of a random choice of the reference channel in relation to a heuristic choice of reference channel on the accuracy of TDOA estimates based on GCC-PHAT. As a measure of accuracy, the difference scores of the computed TDOA values in relation to the approximated ground truth values were computed. Ground truth values are defined as the expected TDOA in samples given the a priori approximated distances between microphones. Table 5 depicts the mean and standard deviation of the difference scores for each channel of the 128 samples in the random and heuristic condition of reference channel selection. Results show an increased mean difference score and variance in the random condition in relation to the heuristic condition.

Table 5

Difference scores of the TDOA values in relation to approximated ground truth values for channel alignment based on randomly selected reference channel and reference channel based on heuristic. Displayed are the mean and standard deviations for each channel of the n=128 samples.

	Random Selection		Heuristic	
MIC ID	**Mean**	**SD**	**Mean**	**SD**
01	5.4	10.7	1.0	0.0
02	1.7	1.9	0.0	0.0
03	5.4	10.3	1.4	1.1
04	12.3	28.5	0.0	0.0
05	3.4	1.9	4.0	0.0
06	27.4	36.8	5.2	2.2
16	5.7	10.5	0.0	0.0

3.2. Analysis 2: Reference Channel on MCCC Channel Selection

The second analysis investigated potential follow-up effects of the choice of reference channel during the alignment procedure on the channel selection and ranking based on the MCCC algorithm. A channel ranking based on the MCCC algorithm was computed for all 128 samples with a reference channel randomly chosen during the alignment procedure and a reference channel chosen based on the delay heuristic. The resulting MCCC based channel rankings from worst to best quality channels for both conditions were compared to the rankings based on STOI scores. As previously described, the MCCC algorithm can retain a minimum set of 2 channels during the selection procedure. These two channels are ranked as the signals of best quality according to the MCCC algorithm. For the random selection of reference channel condition, the number of samples in which the set of the two selected channels was identical to the set of the two best quality channels based on the STOI scores was decreased with n=33 samples compared to the heuristic condition with n=116 samples. As indicated in table 4, STOI scores decreased with increasing distance to the foreground speaker box indicating that STOI score rankings adequately captured the effects of signal attenuation and reverb. The supplementary material shows the channel rankings for the first 10 samples in the random and heuristic condition based on the MCCC algorithm along with the reference channel used for each sample and condition.

4. Discussion

Present findings reveal that the choice of reference channel for alignment effects TDOA accuracy based on GCC-PHAT. More specifically, random selection of the reference channel was associated with increased deviation from ground truth values as well as with increased between sample variability of TDOA estimates. In addition, the choice of reference channel for the alignment procedure affected results of the subsequent channel selection approach based on MCCCs. Selection of the reference channel based on a delay heuristic yielded channel selection results that were congruent with STOI scores for the majority of the utterances. In contrast, random selection of a reference channel was associated with only 26% of the samples in line with STOI scores. Findings suggest that deviations from the ground truth in the alignment procedure as well as the selected reference channel per se might affect subsequent spatial covariance computations involved in the MCCC channel ranking approach and thus yield selection results that are not optimal for subsequent speech recognition steps in terms of speech quality.

In contrast to previous research that used a linear microphone array with N=64 microphones and equal inter-microphone spacings of 2 cm [6, 18], the present research employed only a small subset of microphones and these were distributed with inter-microphone distances up to 5 meters. Therefore it is to be expected that channel differences between microphones are more pronounced in the present data set due increased inter-microphone effects of reverb and sound attenuation. Consequently, when optimizing alignment to a remote channel with substantial reverb effects and a decayed source signal, the spatial correlation of channels with similar reverb shaded degradations could be enhanced and thus confound overall results of the channel ranking.

Notably, the microphones included in the present work were not of the same kind, but differed in terms of their operating principle: dynamic microphones as well as condenser microphones were included in the analysis. Although the present number of channels is very limited, findings show a trend towards increased variability in TDOA estimates in microphones not only as a function of distance to the sound source but also as a function of microphone type in the random selection condition. The condenser microphones were associated with increased variability, particularly when they were located more remotely from the sound source. It remains up to future research to further investigate these tentative findings and also if subband calculations of GCC-PHAT may decrease these effects [11].

Recent research focusing on far field speech recognition in noisy and reverberant conditions with coarsely distributed microphone arrays has focused increased attention on the choice of reference channel. Maximization of cross-correlation coefficients [12] as well as attention-based approaches [17, 21] have been suggested as strategies for reference channel selection. This is in line with the present findings, implying that the choice of reference channel should not be left to chance in environments where microphones are more widely distributed and thus record signals that differ more profoundly with regard to reverb and attenuation.

4.1. Limitations

The spatial accuracy of present calculations was limited by the distance information provided by the VOiCES corpus documentation of the corresponding website [20]. Inter-microphone distances were broadly approximated by building the difference scores between the given distance information of each microphone to the foreground speaker box. Consequently, differences in height were not adequately represented in the derived ground truth distances. In addition, distances were indicated in inches without decimal points which also limits the accuracy of the present conversions to centimeters. Consequently, the present TDOA results of channel alignment were compared to broadly approximated ground truth values. Despite this limitation, present results on channel alignment are meaningful in that they do not only show an increased deviation from ground truth values when selecting a reference channel randomly, but they also show an increased variability in this deviation as compared to a heuristic selection of a reference channel.

Speech recordings of the present data were based on prerecorded LibriVox utterances played by a speaker box positioned in the room. This needs to be taken into consideration as spectrograms between

recorded speech and real human speakers may differ depending on the recording conditions and thus may be distinguishable based on spectral features.

The number of microphones included in the present calculations was constrained to unobstructed microphones that were in line with the speaker box yielding a maximal time delay between subsequent microphones. Furthermore, we did not report the results of a third operating type of microphone included in the corpus, so called MEM microphones. Initial results with MEM microphone recordings used in the corpus could not be related back to the approximated ground truth values and we therefore did not include them in the present paper. This was confirmed by written correspondence with one of the authors of the VOiCES corpus stating that some of the MEM micrphones had a short delay prior to the signal output.

Finally, the MCCC algorithm was introduced as a channel selection method with suggested decreased computational complexity compared to ASR based channel selection approaches [6]. The applicability in real-time settings and computational efficiency of this method when combined with a heuristic for reference channel selection still remains to be explored.

5. Acknowledgements

This research was funded by the Federal Ministry of Education and Research (Germany).

6. References

[1] H. Riaz, M. Stiles, C. Armstrong, A. Chadwick, H. Lee, G. Kearney, Multichannel microphone array recording for popular music production in virtual reality, in: Proceedings of the AES 143rd Convention, New York, NY, 2017, Article Number: eBrief384.

[2] A. Farina, A. Capra, L. Chiesi, L. Scopece, A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production, in: Proceedings of the AES 40th Conference, Tokyo, Japan, 2010.

[3] R. Martin, Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction, in: M. Brandstein, D. Ward (eds), Microphone Arrays, Digital Signal Processing, Springer, Berlin, Heidelberg, 2001, pp. 255–279. https://doi.org/10.1007/978-3-662-04619-7_12.

[4] N. K. Chaudhary, S. Verma, A. Aditya, Sound Source Localization using GCC-PHAT with TDOA Estimation, Journal of Basic and Applied Engineering Research, 1.11 (2014): 54–58. Online ISSN: 2350-0255.

[5] R. Lee, M. Kang, B. Kim, K. Park, S. Q. Lee and H. Park, Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments, IEEE Access 8 (2020) 7373–7382. doi: 10.1109/ACCESS.2019.2963768.

[6] K. Kumatani, J. McDonough, J. F. Lehman, B. Raj, Channel selection based on multichannel cross-correlation coefficients for distant speech recognition, in: Joint Workshop on Hands-free Speech Communication and Microphone Arrays, Edinburgh, UK, 2011, pp.1-6. doi: 10.1109/HSCMA.2011.5942398.

[7] Z.-Q. Wang, D. Wang, All-Neural Multi-Channel Speech Enhancement, in: Proc. Interspeech 2018, Hyderabad, 2018, pp. 3234-3238. doi: 10.21437/Interspeech.2018-1664

[8] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech, and Signal Processing, 24.4 (1976): 320-327. doi: 10.1109/TASSP.1976.1162830.

[9] W. Li, Y. Zhang, P. Zhang, and F. Ge, Multichannel ASR with Knowledge Distillation and Generalized Cross Correlation Feature, in: 2018 IEEE Spoken Language Technology Workshop, Athens, Greece, 2018, pp.463-469.

[10] C. Zhang, D. Florencio, and Z. Zhang, Why does PHAT work well in lownoise, reverberative environments?, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008, pp. 2565-2568. doi: 10.1109/ICASSP.2008.4518172.

[11] M. Cobos, F. Antonacci, L. Comanducci, A. Sarti, Frequency-Sliding Generalized Cross-Correlation: A Sub-Band Time Delay Estimation Approach, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28 (2020) 1270–1281. doi: 10.1109/TASLP.2020.2983589.
[12] J. Dennis, T. H. Dat, Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASpIRE challenge, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Arizona, USA, 2015, pp. 518-524.
[13] M. Wölfel, C. Fuegen, S. Ikbal, J. McDonough, Multi-source far-distance microphone selection and combination for automatic transcription of lectures, in: Proc. Interspeech 2006, Pittsburgh, USA, 2006, paper 1253-Mon2BuP5.
[14] M. Wölfel, Channel selection by class separability measures for automatic transcriptions on distant microphones, in: Proc. Interspeech, 2007, Antwerp, Belgium, 2007, pp. 582-585.
[15] C. G. Flores, G. Tryfou, M. Omologo, Cepstral distance based channel selection for distant speech recognition, Computer Speech & Language, 47 (2018) 314–332. doi: 10.1016/j.csl.2017.08.003.
[16] F. Xiong, J. Zhang, B. Meyer, H. Christensen, J. Barker, Channel Selection using Neural Network Posterior Probability for Speech Recognition with Distributed Microphone Arrays in Everyday Environments, in: Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments, Hyderabad, 2018, 19-24. doi: 10.21437/CHiME.2018-5.
[17] T. Ochiai, S. Watanabe, T. Hori, J. Hershey, Multichannel End-to-end Speech Recognition, 2017. URL: https://www.merl.com/publications/docs/TR2017-035.pdf.
[18] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, I.Tashev, Microphone array processing for distant speech recognition: Towards real-world deployment, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 2012, pp. 1-10.
[19] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech, IEEE Transactions on Audio, Speech, and Language Processing, 19.7 (2011): 2125–2136.
[20] C. Richey, M.A. Barrios, Z. Armstrong., C. Bartels, H. Franco, M. Graciarena, A. Lawson, M.K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, K. Ni, Voices Obscured in Complex Environmental Settings (VOiCES) Corpus, in: Proc. Interspeech 2018, Hyderabad, 2018, 1566-1570. doi: 10.21437/Interspeech.2018-1454.
[21] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, Multi-channel Attention for End-to-End Speech Recognition, in: Proc. Interspeech 2018, Hyderabad, 2018, 17-21, doi: 10.21437/Interspeech.2018-1301.

7. Supplementary Material

Table S1
Channel rankings based on MCCCs for condition with heuristic selection of reference channel for alignment. Channels are sorted upwards from worst quality to the two best quality channels based on MCCC. Columns 6 and 7 denote the two channels that were ranked as best quality channels based on the MCCC algorithm.

	Channel ranking: Heuristic Reference Channel							Channel Alignment
	Rank 1	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**	**Rank 6**	**Rank 7**	**Reference Channel**
Sample 1	16	4	3	6	5	1	2	2
Sample 2	16	5	6	4	3	1	2	2
Sample 3	3	4	16	6	5	1	2	2
Sample 4	16	4	3	6	5	1	2	2
Sample 5	16	4	3	6	5	1	2	2
Sample 6	16	4	3	6	5	1	2	2
Sample 7	16	6	5	4	3	1	2	2
Sample 8	16	5	6	4	3	1	2	2
Sample 9	16	4	3	6	5	1	2	2
Sample 10	3	4	16	6	5	1	2	2

Table S2

Channel rankings based on MCCCs for condition with random selection of reference channel for alignment. Channels are sorted upwards from worst quality to the two best quality channels based on MCCC. Columns 6 and 7 denote the two channels that were ranked as best quality channels based on the MCCC algorithm.

	Channel ranking: Random Reference Channel							Channel Alignment
	Rank 1	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**	**Rank 6**	**Rank 7**	**Reference Channel**
Sample 1	4	2	3	16	1	5	6	6
Sample 2	5	6	16	4	3	1	2	1
Sample 3	2	16	1	6	5	3	4	5
Sample 4	4	2	3	16	1	5	6	6
Sample 5	2	6	5	16	1	3	4	4
Sample 6	16	4	3	6	5	1	2	2
Sample 7	16	6	5	4	3	1	2	2
Sample 8	2	16	1	4	3	5	6	6
Sample 9	16	4	3	6	5	1	2	2
Sample 10	6	5	16	4	3	1	2	1