IMPROVED SIMULATION OF REALISTICALLY-SPATIALISED SIMULTANEOUS SPEECH USING MULTI-CAMERA ANALYSIS IN THE CHIME-5 DATASET

Jack Deadman, Jon Barker

Department of Computer Science, University of Sheffield, UK
{jdeadman1, j.p.barker}@sheffield.ac.uk

ABSTRACT

Room simulation is an essential tool in the development of distant microphone ASR and source separation. However, most commonly used simulated datasets adopt uninformed and potentially unrealistic speaker location distributions. In earlier work, we analysed a 50-hour audio-visual dataset of multiparty recordings made in real homes to estimate typical angular separations between speakers. We now refine and extend this work using a multi-camera analysis to estimate full 2-D speaker location distributions. Results show that commonly used simulated datasets use unrealistically large angular separations, but unrealistically small ranges for target to interferer distance ratios. We generate more realistically distributed datasets and use them to re-evaluate state-of-the-art source separation and ASR approaches. Results suggest that imposing realistic angular separation distributions makes datasets more challenging, however, the pattern when using realistic distance ratios is more complicated and can depend on room size.

***Index Terms*—** automatic speech recognition, data simulation, source separation

1. INTRODUCTION

Acoustic room simulation [1, 2, 3] is an essential tool for developing distant microphone automatic speech recognition (ASR) systems. Simulation allows for clean reference signals to be used in evaluating speech enhancement [4, 5], for arbitrary large training data [6] to be constructed and for targeted evaluation and analysis of the performance of ASR systems [7]. Simulation is commonly used for generating *training data*. For example, for augmenting real training data, or for providing ground truth information when training supervised speech enhancement systems. For the simulated data to be useful, it needs to match the distribution of the real target data [8]. However, simulation is also often used for generating *evaluation data*. In such cases, the need for realism is even more crucial: a poor simulation can result in wasted effort, i.e., by promoting approaches that work in simulation but not in real situations.

Although modern methods for acoustic room simulation can accurately model the physics of sound propagation, e.g., [3], this is only one part of the problem. Room simulations are driven by their metadata, e.g., the room size, location of sources, T_{60} time and so on. The distribution of this metadata needs to be carefully considered. If it is poorly motivated the resulting dataset can overemphasise the importance of one component of a speech processing system over another. For example, the importance of beamforming approaches can be overplayed if simulations have unrealistically large angular separations between speakers [9].

In our work, we are interested in simulating multiparty conversations for distant microphone speech recognition research. Previously, we used a large real audio-visual dataset (CHiME-5 [10]) to look at one aspect of this problem, angular speaker separation [11]. Our methodology was to use camera data from single devices to estimate and hence simulate realistic angles between overlapping speakers. In this paper, we extend this analysis by using multiple cameras. This allows the 2-D room location of target and interference speakers to be estimated. This data can then be used to correctly simulate the full spatial distribution of speakers, and hence produce data with realistic speaker properties such as signal-to-noise ratio (SNR), angular separation and direct-to-reverberant energy ratio (DRR).

The paper is organised as follows. Section 2 reviews previous simulated spatialised-speech datasets and their role in automatic speech recognition research. In Section 3, we introduce our methodology for calculating speaker locations using multiple devices from annotated single-device data. We then use these positions in Section 4 to estimate the relative distance of speakers and interferers in the CHiME-5 datasets. This analysis is used to inform a simulation with an improved estimate of angular separation and speaker distance.

2. BACKGROUND

The speech enhancement and source separation fields rely heavily on simulated datasets constructed by convolving room impulse response (RIRs) with clean utterances, for example from WSJ [12] and LibriSpeech [13]. The spatialised version of WSJ0-2MIX was introduced in [14], which became a common benchmark for multichannel source separation algorithms. In recent years, deep learning techniques have performed so well in these scenarios that more challenging datasets have been required. WHAM! [15] increased the challenge by adding real background noise and then WHAMR! [16] extended WHAM! by using *reverberant* noisy mixtures. Both multichannel WSJ0-2mix and WHAM! use the WSJ corpus [12] as their source for clean speech signals, and both randomise speaker positions uniformly in the room. LibriMix [8] was introduced to compliment WHAM! with different source material but no spatialised version has been created. Finally, SMS-WSJ[17] was introduced to address the fact that these datasets provide a small amount of training data compared to what is required for acoustic modeling.

Attempts have been made towards creating simulations that mimic more realistic temporal overlap in simulation [18, 19]. However, little progress has been made towards generating data-driven speaker positioning in these setups [11], even though there is a wealth of behavioural research showing that people in multi-party conversations observe social rules that govern how they are spaced, i.e., the field of proxemics [20]. Due to these spatial mismatches, amongst others, deep learning techniques may perform well in simulated environments, but then perform poorly in real domestic scenarios [21].

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

3. METHODOLOGY

Realistic speaker location distributions are learnt from the CHiME-5 dataset, a unique dataset that contains long unscripted recordings of informal 4-person ‘parties’ recorded across many homes. Analysing CHiME-5 allows us to gain insight into the natural behaviour of people in conversational settings. The data comprises recordings from Microsoft Kinect v2 devices placed unobtrusively at the edges of rooms. The devices contain a microphone array with an integrated camera. The video recordings, which have overlapping fields of view, allow speaker location to be estimated.

In order to accurately estimate the position of speakers in the room, several challenges need to be addressed. First, accurate speaker locations need to be estimated in the image-space of each of the devices (Section 3.1). Second, in order to map from the device image spaces to the physical room space, the location and orientation of each of the devices needs to be known. This step uses a calibration procedure that estimate actual location of the devices given initial rough sketch estimates provided with CHiME-5 (Section 3.2). Finally, a procedure for mapping into room space is required that is robust to errors in the video annotation and camera parameter estimation (Section 3.3).

3.1. Speaker location annotation

An annotation tool[1] is used that employs a mixture of optical flow tracking and manual guidance to allow an annotator to efficiently and accurately track the location of each person’s mouth (or estimated location in case of occlusion). Annotations are made at 100 ms intervals with occasional dropped frames in-filled via linear interpolation. Annotated tracks are reviewed and corrected as necessary.

The CHiME recordings are each around 150 minutes in duration. Each is composed of three separate phases of roughly equal length, focusing on activity in different areas of the living space (kitchen, dining, living room). A sample of the data is used composed of 5 minute segments from the middle of each phase. There are 20 separate CHiME party recordings and so 60 5-minute segments, each recorded with 5 or 6 devices, to make a grand total of 342 videos of which 186 were annotated containing participants. The video may feature between 0 and 4 participants depending on the party phase and the device location. Note that many of the environments are ‘open plan’ flats, so devices located in a living room, or kitchen area, can detect participants in the dining area, for example.

3.2. Camera calibration

In the CHiME-5 dataset, the floorplan of each of the rooms is provided through sketches. These sketches include the walls and their measured length and rough locations of the devices and their rotations. This provides a good initial starting point for the true location of the devices, but if they are used naïvely, the final estimate of position estimates will be poor.

To address this issue, the devices are calibrated using an optimisation procedure. If three cameras detect the same person, then the vectors produced from their observation angle should intersect at the same point. This is formulated by minimising the following objective function,

$$J(\Theta) = \sum_{k=1}^{K} \frac{1}{L^{(k)}} \sum_{l=1}^{L^{(k)}} ||x_l^{(k)} - c_k||, \quad (1)$$

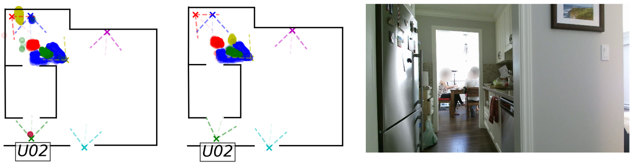

Fig. 1. Results from running the calibration process. The image shows the process has successfully calculated that the device U02 should be rotated in the floorplan. This calibration process has resulted in estimates of positions which are more plausible.

where K is the number of samples, $L^{(k)}$ is the number of intersections for sample k, $x^{(L)}$ is a point of intersection, and c is the centre point of the intersections, i.e., the objective is to minimise the distance between the intersections and the mean intersection point. Therefore, an optimal solution places the device such that all the devices “agree” with each other. We minimise this objective function using stochastic gradient descent. The parameters are the x,y coordinates and rotation of each of the devices (18 parameters in total for a session). A sample in this formulation is a vector containing the detected angle in each of the devices. If a camera does not detect the person, the computed gradient is set to zero for the corresponding parameters for that device.

The calibration procedure assumes that the devices remain stationary. The devices are supposed to remain stationary throughout a session. Although the devices are at fixed locations, analysis of the data indicates that small movement occasionally occur, presumably when they have been accidentally disturbed by participants. For each device three parameters are being estimated (x, y, yaw) and three ignored (z, pitch and roll).

An example of running this calibration procedure is shown in Fig. 1. The figure shows how the calibration procedure has correctly adjusted device U02 to more slightly rotated clockwise. By then examining the example frame from the video, we can see this is a sensible adjustment as the camera in the floorplan has been slightly rotated matching what we can see in the video i.e., the initial floorplan has the device facing completely forwards whilst in fact it is rotated slightly to the right. Similar assessments were made for the other sessions to verify the calibration was working. However, the technique requires at least three cameras to see a person so it is not always possible to update a camera position.

3.3. Estimating speaker location

After camera calibration, for each frame, we estimate a person’s true angle to the device given the annotated observations in the images. We model this using a Gaussian distribution with the mean set to the angle in the annotation,

$$\theta \sim \mathcal{N}(\mu = d, \sigma^2) \quad (2)$$

The variance, σ^2, models inaccuracies in the annotation and in estimation of the device angle. This parameter has been set empirically and is tuned to 10 degrees in the experiments that follow.

Given these annotations in isolated cameras, we can estimate the most probable location of speakers in a two-dimensional space. Given a 2D position in the room, an angle can be computed in each of the devices by projecting the position (x, y) into an angle for each of the devices where r_i is the rotation of the device i,

$$wrap(\theta) = \angle exp(j\theta) \quad (3)$$

[1] The tool is available to use, https://github.com/jackdeadman/tracking-annotator

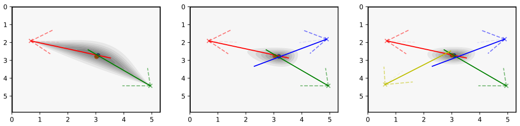

Fig. 2. Illustration of Eq. 5 showing how adding more cameras changes the estimate of positions. The darker areas indicate a higher probability of the person being in that location given the detections in each of the cameras.

$$project_i(x, y) = wrap(atan2(y, x) - r_i) \tag{4}$$

These observation angles can then be combined to give a probability of a position given the annotations $o_1, ..., o_N$, where N is the number of devices,

$$P(x, y|o_1, ..., o_N) = \prod_{i=1}^{N} \mathcal{N}\left(project_i(x, y); \mu = o_i, \sigma^2\right) \tag{5}$$

If a speaker is not detected in a camera then the probability mass is distributed uniformly across all angles.

The probability function in Eq. 5 is depicted in Fig. 2. In the figure an artificial setup is created with the true location of the person placed at the position (3, 3). An error has been artificially added to the simulated detections to show how the probabilities of the position change as more devices are added to the formula. Even though the final camera provides a poor estimate of the position, it does not skew the distribution away from the true location.

Given this probability distribution, a location can be estimated by either choosing the peak,

$$pos_{max} = \underset{x \in W, y \in H}{\arg\max}\ P(x, y) \tag{6}$$

or by computing the expected value,

$$pos_{exp} = \left[\sum_{x \in W} xP_X(x), \sum_{y \in H} yP_Y(y)\right]^T \tag{7}$$

The max point provides the most plausible estimates when the devices are well calibrated and close to each other. In the more difficult cases, i.e., devices facing each other, the expected point resulted in more plausible estimates and the max point was found to be very sensitive to small changes in the image-space location estimates. We use pos_{exp} as the estimate of the speaker position.

4. ANALYSIS

4.1. Estimating angular separation using 2-D positions

Using the procedure described in the methodology, the 2-D positions of people in the dataset are computed. This allows us to refine the estimate of angular separation we previously reported. The automatic pose detection method used in [11] was limited by the field of view of the device, i.e. separation of more than 84.1 degrees is not possible as at least one of the speakers would not be visible. The automatic method also underestimates extremely small angles due to the speakers occluding each other. Now that the 2-D positions of speakers have been estimated, we can project these positions onto the devices to get the angle they would be if the device could see them. This is important as even if the participant is not visible, their speech would still be recorded. The plot in Fig. 3. shows the updated angular separation from projecting the positions into a reference device. The reference device is chosen by selecting the device which on average throughout a segment people are closest to, this device is then constant for the entire segment. The plot shows that the angle is still very narrow between the speakers but not as extreme as first thought. The plot also shows the separation of the speakers if the labels are used directly, which shows a similar distribution as to what was previously reported. The angles from projection show us the first estimated distribution was accurate but with a longer tail.

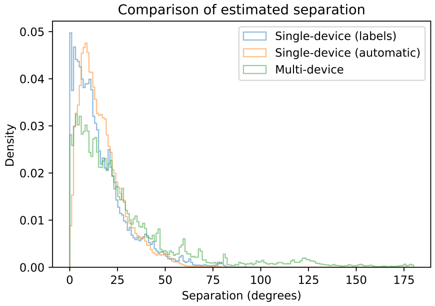

Fig. 3. Distribution of the angular separation estimates for different estimation approaches. Shown are two single-device approaches (one automatic and one using labelled data) and a multi-device approach which uses a combination of cameras to produces 2-D position estimates, which are then projected into the reference device.

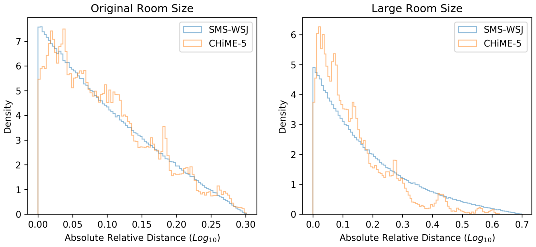

Fig. 4. Comparison of the absolute log of the ratio between speaker and competing speaker. Under the constraint speakers are between 1 and 2 metres (left) from the device, speakers position themselves somewhat randomly. In a larger room (right) setting they position themselves closer to each-other i.e., form a group.

4.2. Estimating relative distance

These positions now allow us to estimate the distance speakers are away from the microphone with respect to a competing speaker. For a frame in the video that has two or more people with estimated positions, two random people are selected and one is assigned to be the target speaker. The plot in Fig. 4 compares the absolute value of the log ratio of the target speaker and interferer, i.e., $|log_{10}(D)|$,

$$D = \frac{d_{target}}{d_{interferer}} \tag{8}$$

where d_{target} and $d_{interferer}$ are the distances away from the device for the target speaker and interferer respectively.

Here we can see that, on average, people stand closer together in this social setting as compared to positioning randomly. This is an expected result but shows a data driven approach to showing the phenomenon. The plot on the left shows the relative distance of speakers

Table 1. Results from the cACGMM baseline system comparing several datasets with *fit* (F) and an *uninformed* (U) distributions for angular separation (Φ) and relative distance (D).

	Name	Φ	D	PESQ	STOI	SDR	WER
Original	SMS-WSJ [17]	U	U	2.07	0.82	12.35	18.25
	Prev. work [11]	F*	U	1.85	0.74	9.0	31.49
	O+Φ	F	U	1.91	0.76	9.80	28.25
	O+D	U	F	2.06	0.82	12.17	18.49
	O+Φ+D	F	F	1.90	0.76	9.79	28.09
Large	L-SMS	U	U	2.01	0.78	11.38	22.49
	L+D	U	F	2.04	0.80	11.59	21.73
	L+Φ	F	U	1.84	0.73	8.40	34.21
	L+Φ+D	F	F	1.83	0.73	8.09	36.07

**Angular separation distribution fit on different data.*

when we constrain the estimates to be in the range of positions that can be found in SMS-WSJ. In the plot on the right, the distribution in SMS-WSJ is extended to between 1 and 5 metres and the range in CHiME-5 has been matched. From this plot we can see that the distribution is random when looking at a small room. But when looking at a larger room, people tend to gather in groups.

5. REALISTIC SPEAKER LOCATION IN SIMULATION

To evaluate the impact that speaker positioning has on source separation and ASR we create a series of datasets that use data driven approaches to create more realistic setups. We run ASR experiments to show the impact that these setups have on performance of source separation and recognition to illustrate the potential impact of the mismatch between typical simulation and real data.

5.1. Experimental setup

Experiments use the baseline system described in [17], namely, a cACGMM mask estimator is used with a minimum variance distortionless response (MVDR) beamformer and a factorised time-delayed neural network (TDNN-F) based acoustic model. The experiments measure how the baseline performance changes when the SMS-WSJ dataset enforces realistic speaker distributions. We compare the impact of the relative distance distribution and the updated angular separation distribution. To account for the dependency on the change in relative distance distribution when the absolute distance is larger, an additional set of datasets are created with a room size mean set to (12, 8) which samples speaker distances between 1 and 5 metres, which we name *large*. The *original* dataset has a mean room size of (8, 6) and samples distances between 1 and 2 metres.

For realistic separation, the angular separation of the speakers are sampled from a Gaussian kernel density estimate of the angular separation distribution from multi-device as shown in Fig. 3. For realistic relative distance, the first speaker's absolute distance is sampled in the same way as SMS-WSJ. The competing speaker's distance is then drawn by sampling from a conditional distribution. The conditional distribution is computed by first modelling the joint distribution of absolute distance and relative distance, again using a Gaussian kernel density estimate.

5.2. Results

The results in Table 1 show that the updated angular separation still has a large impact on the performance of ASR and speech enhancement, but slightly less extreme than first reported in our prior work. The impact of fitting the relative distances shows a more complex relationship. In the original room size of SMS-WSJ the performance of the system decreases when the relative distance distribution is enforced. The performance however increases when this is combined with the angular separation. When extended to a large room the results show that a realistic distance actually results in a easier dataset. In the largest room using the fitted distribution for angular separation and relative distance produces the most challenging dataset.

Table 2. Enhancement and ASR performances when using MVDR with estimated masks (cACGMM), oracle masks (IBM), or directly using pre-mixed signals (Image) for large rooms under various speaker spatial distributions: baseline (L+SMS), baseline plus realistic distances (L+D), plus realistic angular separation (L+Φ), or both (L+D+Φ).

	Mask	Enh	PESQ	STOI	SDR	WER
L-SMS	cACGMM	MVDR	2.01	0.78	11.38	22.49
	IBM	MVDR	2.01	0.80	12.10	16.68
	Image		2.00	0.80	13.20	9.66
L+D	cACGMM	MVDR	2.04	0.80	11.59	21.73
	IBM	MVDR	2.05	0.81	12.36	16.41
	Image		2.03	0.81	13.53	9.63
L+Φ	cACGMM	MVDR	1.84	0.73	8.40	34.21
	IBM	MVDR	1.88	0.76	10.23	21.81
	Image		2.00	0.80	13.24	9.80
L+Φ+D	cACGMM	MVDR	1.83	0.73	8.09	36.07
	IBM	MVDR	1.88	0.76	10.13	21.96
	Image		2.02	0.81	13.42	9.45

In Table 2 we show that the performance of the mask estimator (cACGMM) is consistent with the performance using an oracle ideal binary mask (IBM). Looking at the performance of the system on the Images, i.e., the signal after convolution but before mixing, we can see that these raw signals were easier for the ASR system to transcribe in L+Φ+D (WER 9.45%) compared with L+Φ (WER 9.80%) but this performance is reversed after mixing. This shows it is not the absolute distance to the microphone making the dataset more difficult but it is indeed the relative distance.

6. CONCLUSIONS

In this paper we have contributed an analysis of the relative distance between speakers and competing speakers in unscripted dinner parties. We have detailed our methodology and the challenges involved in deriving this estimate. The analysis also contributed an updated estimate of the angular separation of speakers in the CHiME-5 dataset. Our experimental work shows the relationship of angular separation and the challenges it produces when the angle narrows. Our work has also demonstrated the complicated relationship of relative distance and its affect on performance.

We have produced speaker location labels for a subset of CHiME-5, which will be released alongside this work[2], allowing for our analysis to be reproduced. Derived 2-D positions are also released alongside this, to allow for further analysis. In addition to this we release RIRs and metadata for the datasets produced, which we hope can be used as additional benchmarks, i.e., allowing the community to analyse the performance of their ASR systems with respect to angular separation and relative microphone distance in a comparable manner.

[2] `https://chime.jackdeadman.com`

7. REFERENCES

[1] Jont B Allen and David 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.

[2] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[3] Dirk Schröder and Michael Vorländer, "RAVEN: a real-time framework for the auralization of interactive virtual environments," in *Forum acusticum*. Aalborg Denmark, 2011, pp. 1541–1546.

[4] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "BSS_EVAL toolbox user guide–revision 2.0," 2005.

[5] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[6] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[7] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[8] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[9] Chao Pan, Jingdong Chen, and Jacob Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, 2014.

[10] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[11] Jack Deadman and Jon Barker, "Simulating Realistically-Spatialised Simultaneous Speech Using Video-Driven Speaker Detection and the CHiME-5 Dataset," in *Proc. Interspeech 2020*, 2020, pp. 349–353.

[12] Douglas B Paul and Janet Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[15] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[16] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 696–700.

[17] Lukas Drude, Jens Heitkaemper, Christoph Boeddeker, and Reinhold Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[18] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[19] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.

[20] Edward T Hall, "A system for the notation of proxemic behavior 1," *American anthropologist*, vol. 65, no. 5, pp. 1003–1026, 1963.

[21] Matthew Maciejewski, Gregory Sell, Yusuke Fujita, Leibny Paola Garcia-Perera, Shinji Watanabe, and Sanjeev Khudanpur, "Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 165–169.