Simulating realistically-spatialised simultaneous speech using video-driven speaker detection and the CHiME-5 dataset

Jack Deadman, Jon Barker

Department of Computer Science, University of Sheffield, UK

jdeadman1@sheffield.ac.uk, j.p.barker@sheffield.ac.uk

Abstract

Simulated data plays a crucial role in the development and evaluation of novel distant microphone ASR techniques. However, the commonly used simulated datasets adopt uninformed and potentially unrealistic speaker location distributions. We wish to generate more realistic simulations driven by recorded human behaviour. By using devices with a paired microphone array and camera, we analyse unscripted dinner party scenarios (CHiME-5) to estimate the distribution of speaker separation in a realistic setting. We deploy face-detection, and pose-detection techniques on 114 cameras to automatically locate speakers in 20 dinner party sessions. Our analysis found that on average, the separation between speakers was only 17 degrees. We use this analysis to create datasets with realistic distributions and compare it with commonly used datasets of simulated signals. By changing the position of speakers, we show that the word error rate can increase by over 73.5% relative when using a strong speech enhancement and ASR system.

Index Terms: speech recognition, speech enhancement, data simulation

1. Introduction

It is becoming commonplace to find voice assistants in homes. These devices use an array of microphones to exploit spatial cues to enhance speech in the desired direction whilst suppressing competing sounds such as noise and competing speakers in other directions. It is well known that these systems perform better when there is a greater angular separation between speakers. However, studies in the behaviour of people have shown that in a social setting, people tend to stand close to each-other [1]. We will explore this conflict by analysing the behaviour of people in social settings and the impact it has on current speech enhancement techniques and automatic speech recognition (ASR). Knowing the true behaviour of speakers will help in understanding how best to design future microphone array algorithms and hardware.

To benchmark speech enhancement techniques, a controlled environment is required where a version of the audio before distortion is available. This is typically achieved using databases of simulated signals, which are created by generating room impulse responses (RIR) through simulation, e.g., the image method [2], and then convolving the RIR with the clean audio. Simulating the complexities of the real world is an incredibly difficult task but an important gap that needs to be bridged to provide meaningful results before algorithms are tested on real data. Advances have been made in improving the realism of simulations [3], for example, by simulating non-cuboidal room shapes [4]. However, the distribution of speaker location is an aspect that has been largely overlooked.

Often in multi-channel speech enhancement, when reporting results, an overall performance is presented with no information on how the performance relates to the speaker separation distribution in the dataset being used. There are many important discussions to have when evaluating the performance of speech enhancement systems such as the performance metric that should be used [5]. We believe the separation of speakers is an important topic that should also be discussed when evaluating these systems. This paper argues that current simulated datasets such as [6, 7, 8] do not represent the spatial separation of speakers in typical social settings and therefore, may produce overpromising results. For this study, the video data captured during the recording of the CHiME-5 dataset is used for the analysis of speaker separation. We will use analysis from cameras capturing videos from the perspective of 114 microphone arrays recording 50 hours of social interaction in 20 homes. For a full description, see [9].

The paper is organised as follows. Section 2 describes how person location information was automatically extracted from the videos, and the evaluation used to establish confidence in the data. Then, in Section 3, this data is used to estimate the real distribution of angular separation of speakers in CHiME-5 and finally, in Section 4, we explore the impact that enforcing this realistic distribution has on speech enhancement and ASR.

2. Azimuth angle of speakers

The CHiME-5 dataset consists of 20 dinner party sessions, with each party broken into three stages: cooking, dining and after-dinner socialising. Each of these stages typically takes place in different rooms of the house, i.e., *Kitchen*, *Dining*, *Living* rooms respectively. A room is captured by two Microsoft Kinect V2 devices, consisting of a 4-channel linear microphone array and a 1080p camera. The location of the devices was chosen such that they were not obstructing the participants, i.e., at the edge of the room looking into the party. This means the placement of the devices do not necessarily maximise the separation of speakers but more closely mimic the placement of a device in a real home use case.

To find the angle of the speakers, we will map the position of speakers in the image to an azimuth angle. The azimuth is the target because like most linear arrays, the Kinect is linear in the horizontal plane, as this is where most spatial diversity occurs. This means for our analysis, the x pixel of a speaker in the image is the most important feature to capture. The angle of azimuth can be approximated from the x pixel using

$$pixel2angle(x) = \frac{x \times 84.1}{1920} \quad (1)$$

where x is the x pixel index, 84.1 is the field of view of the camera in degrees, and 1920 is the resolution of the video.

To detect speakers, two different 'out-of-the-box' tools were used, the Dlib CNN face detector [10] *(face)* and the OpenPose keypoint detection library [11] *(pose)*. These tools can

both be regarded as state of the art but have different strengths and weaknesses. The face detection system is only able to locate a person if they are facing the camera or their profile view is visible. The pose detection system is able to locate people turned away from the camera but suffers from more false detections. These detection systems were run on each of the frames in the 114 videos in isolation.

2.1. Creating hand-annotated ground truth

To establish the confidence in these tools, two different ground truth annotations were created. First, isolated frames were annotated by sampling every five minutes in all the videos in the development set and then annotating a bounding box around the speaker's face as well as annotating the position of the mouth[1]. Second, annotations were created by a separate tool which allows an annotator to use a mouse to track a person as they move by following their mouth position. This tool allowed for far more data to be annotated at the expense of reduced accuracy. This process was repeated for all four speakers and three cameras in one session of the development set.

2.2. Evaluation of detection

Performance of the detection systems was evaluated using two evaluation metrics, which both use the frame annotations.

1. How well does it find people? *(Detection)*
2. When it detects a person, how accurate are the detections? *(Accuracy)*

Detection For *face*, a detection is considered correct if the intersection over union score is above 30%. For *pose*, a detection is correct if the nose position is within 53 pixels[2] of the annotated mouth position. The nose is used as this is the closest keypoint to the mouth provided by the pose detection system.

The detections are paired up with the ground-truth such that the maximum score is provided; this is legitimate because we are not evaluating the speaker assignment. The results are shown on the left-hand side of Table 1. The face-detection system finds fewer people than *pose* as it fails when people are facing away from the camera. However, high precision means we can be fairly confident that it is finding true faces. *Pose* complements these errors as it misses fewer people but yields more false-positives. Note, a low recall will not hinder the separation analysis that follows as long as the position of persons missed is at random with respect to screen position (see Section 3).

Accuracy Next, we look at the accuracy of these detections by measuring the horizontal and vertical distance in pixels to the mouth, X and Y (Table 1, rhs). For the face detection system this was estimated as a point on the horizontal centre of the bounding box and 74%[3] down from the top, and for the *pose* detection system the nose keypoint position was used as an estimate of the mouth. The same assignment process is used as before. The results show that both detection systems can locate the mouth with around 1 degree of error (1 degree $\approx$ 23 pixels) on average. (Note, in one device an oddly placed mirror mislead the detection systems leading to many large unrepresentative errors. This effect was not seen in any of the other 113 cameras and so was treated as an outlier and the device was removed from the evaluation.)

[1] https://github.com/jackdeadman/video-annotation-tools

[2] Threshold chosen by fitting a two-component Gaussian mixture model on the paired data.

[3] The ratio which minimises the Y distance.

Table 1: *Results are shown from eight devices in two different sessions. (Pr: Precision, Re: Recall, F1: F-Score). 558 faces have been hand-annotated. Video resolution:* 1920×1080. *Accuracies are mean* $\pm$ *standard error.*

	Detection			**Accuracy (px)**		
	Pr.	Re.	F_1	X	Y	Euclid.
Face	98.7%	36.6%	52.5%	23 ± 2	18 ± 1	32 ± 2
Pose	94.1%	60.5%	72.9%	24 ± 3	27 ± 2	40 ± 4

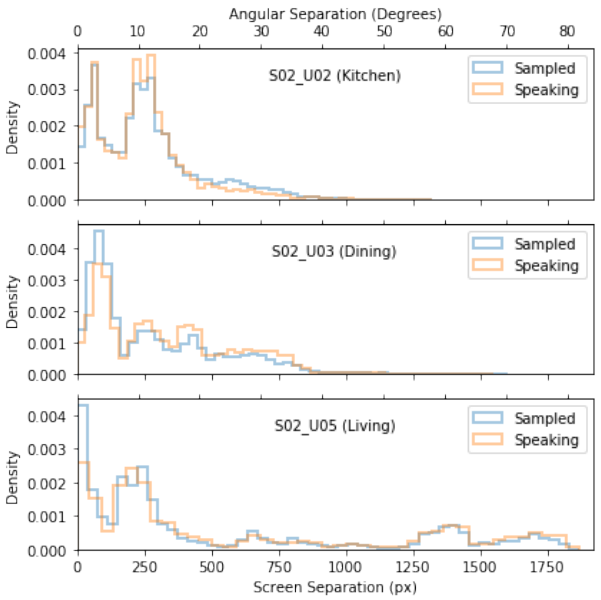

Figure 1: *Comparison of separation distributions when randomly sampling with using active speakers.*

3. Analysis of spatial separation

We now wish to use the angular position estimates from the previous section to estimate angular separation between *active speakers*. The CHiME-5 transcript can be used to recover speaker activity state of identified speakers, but we only know speaker identity for the small annotated video subset. We will, therefore, assume that separation is independent of speaker activity state after first testing this on the annotated subset.

For each video frame in which two or more active speakers are detected by the *face* system, we pick two at random and compute their angular separation. Note, this approach is valid even considering the low recall of the detector assuming the missed detections are missed at random with respect to location. This is then repeated but now sampling pairs of people regardless of speaking activity state. The resulting distributions are compared in Figure 1. The similarity of the distributions suggests that person separation is largely independent of speaking state. This may seem unusual, i.e., people speaking at the same time might be expected to be closer together. However, overlapping speakers may be from competing conversations, and inactive speakers are still 'socially engaged' and therefore standing at conversational distances from each other. The figure also highlights the variety in the distributions between the different devices. The distributions have clear distinct peaks indicating speakers are often in the same locations.

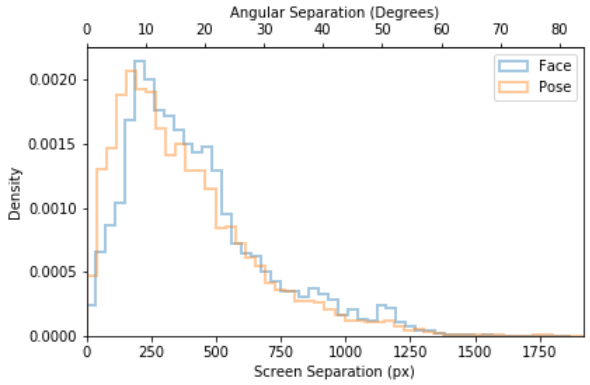

Figure 2: *Comparison of the separation distributions created from the two detection techniques.*

We can now measure person separation across all 114 cameras without regard for speaker activity state and take this as a proxy for overlapped speaker separation. Analysis is repeated with both *face* and *pose* detectors (Figure 2). Even though the two systems have complementary errors, the resulting distributions are similar. Both distributions show that few detections have a separation around 0 pixels. This observation is likely due to the fact the detection systems are not able to detect a person if that person is being occluded by another person, rather than being directly caused by any specific human behaviour.

In Table 2 the overall statistics for the dataset are shown. The mean and standard deviation for the position results are the averages of the mean and standard deviation of each of the sessions. We average over sessions as the initial placement of the device will affect these statistics. Both the detection systems have a slight skew but significant to the left, indicating a bias in the detection systems. Both detection systems show how small the separation angle is between the speakers, with both showing similar separation angles even though the two different approaches have different characteristic.

Table 2: *Position and separation of speakers throughout the dinner parties. The centre of the screen is 0 pixels/degrees. Results are average ± standard deviation.*

	Position		**Separation**	
	Screen (px)	*Angle (°)*	*Screen (px)*	*Angle (°)*
Pose	-23 ± 323	-1 ± 14	380 ± 268	17 ± 12
Face	-35 ± 302	-2 ± 13	427 ± 274	19 ± 12

4. Analysis of simulated datasets

Having estimated the distribution of the speakers throughout CHiME-5, we can compare this with the distribution of simulated datasets. WSJ0-2mix [12] is a commonly-used dataset for source separation, with a spatialised version introduced in [7]. The position of two people is randomly sampled inside a shoebox room with a constraint that speakers cannot be too close to the microphone array and not too close to each other. The latter constraint yields some not immediately obvious consequences. In Figure 3 (left) we compare a proposed dataset where we sample positions according to the distribution we analysed, a recent

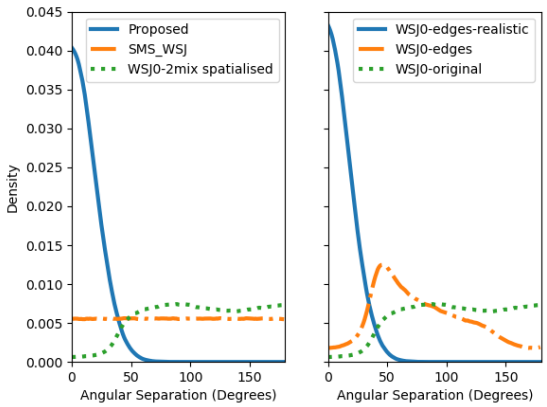

Figure 3: *Comparison of the angular separation in simulated datasets. We compare the datasets SMS_WSJ [6] and WSJ0-2mix spatialised [7] with adapted versions of their setup.*

dataset SMS_WSJ [6] and WSJ0-2mix spatialised. Here we can see the overlaps between the proposed distribution and the two datasets are surprisingly small. The minimum distance constraint in WSJ0-2mix spatialised means that few samples in the dataset have a low separation angle. The proposed dataset was created by exploiting the observation that the speaker positions from the device are normally distributed, therefore to generate the dataset the position standard deviation of 14 degrees was used to approximate the separation distribution indirectly.

To illustrate the importance of reporting the angular separation we, will compare the performance of identical speech separation and recognition systems when tested on the same dataset but with only the speaker position distribution changing. We report source separation metrics and word error rate (WER).

4.1. Experimental setup

Experiments use the baseline system described in [6], namely, a complex angular central Gaussian mixture model (cACGMM) [13] mask estimator is used with a Minimum Variance Distortionless Response (MVDR) beamformer and a factorised time-delayed neural network (TDNN-F) based acoustic model. For the first set of experiments measure how the baseline performance changes when the SMS_WSJ dataset enforces a realistic spatial distribution. In the original setup, the target speaker was placed in the room by randomly sampling a distance and an angle from the microphone array, and a competing speaker is placed at a uniformly sampled angular distance. To generate the two speaker positions for the 'realistic' distribution, an angle is uniformly sampled around the array. Using a Gaussian distribution with a standard deviation of 14 degrees and a mean set by that chosen direction, the two speaker directions are sampled. The speaker distances are then chosen by sampling uniformly from 1-2 metres, i.e., the same as SMS_WSJ. The remaining random parameters are identical to SMS_WSJ. This does not necessarily create a realistic setup because in CHiME-5 the arrays were placed at the edge of the room and here they are placed in the centre of the room. However, it does let us see how the performance of the system changes when speakers have the separation that was present in the real data.

The second set of experiments compares the WSJ0-2mix spatialised setup *(wsj0-original)*, with a variation of the setup

Table 3: *Results from changing the positions of people in the SMS_WSJ database. Oracle results are shown in grey.*

Mask	Enh.	Data	SDR	PESQ	STOI	WER
cACGMM	MVDR	Prop.	9.0	1.85	0.74	31.49
cACGMM	MVDR	SMS	12.3	2.07	0.82	18.15
cACGMM	Mask	Prop.	7.1	1.73	0.71	49.09
cACGMM	Mask	SMS	9.5	1.83	0.78	40.01
None	Ch=0	Prop.	-0.4	1.49	0.66	78.93
None	Ch=0	SMS	-0.4	1.50	0.66	78.73
IBM	MVDR	Prop.	10.4	1.88	0.77	21.23
IBM	MVDR	SMS	12.9	2.06	0.83	14.23

where we place the microphones at the edge of the room *(wsj0-edges)*, in both cases speakers are positioned uniformly in the room with constraints on minimum distances. These are then compared with a setup with the microphones at the edges but with the realistic distribution enforced. This uses the same angle generation method as the previous experiment but with a distance sampled from 1-3 metres *(wsj0-edges-realistic)*. The comparison of the distribution created from this setup is shown in Figure 3 (right). Placing the microphones at the edge of the room resulted in a distribution closer to the real data, but the tail is still far larger than that observed in the real data. Note the distribution of this realistic setup is slightly different from the distribution created in the previous setup. This is due to the resampling of points when they are outside of the room.

4.2. Results

The results from changing the placement of sources in SMS_WSJ are shown in Table 3. We can see that by only changing the location of speakers the WER of the cACGMM system increases by over 13% absolute (73.5% relative) when using the MVDR beamformer, this is a system that contains similar components to the best performing systems on the CHiME-5 dataset [14, 15]. If a system can be made that is more robust to smaller separation angles, then there is huge potential to creating an overall better-performing ASR system. The oracle Ideal Binary Mask (IBM) comparison shows that even with perfect knowledge, the beamformer approach performs significantly worse with the new dataset. Multi-channel approaches that do not use a beamformer may offer a solution to this [7, 16]; however, they rely on closely matched training data.

Finally, the results from converting the WSJ0-2mix spatialised setup to be a more realistically distributed dataset are presented in Table 4 using the cACGMM MVDR system. Here we only show source separation results as the baseline acoustic model would be mismatched with the data from this setup, and the dataset does not provide an agreed-upon training set for acoustic model training. Surprisingly, placing the microphones at the edge of the room does not make the dataset anymore challenging than the original setup, the performances are fairly comparable. Once the more realistic distribution is enforced, then we see that the dataset becomes slightly more challenging. Interestingly, the performances are fairly comparable, which is surprising considering the separation differences. Unlike SMS_WSJ, WSJ0-2mix spatialised does not contain any background noise which could explain this similar performance.

Table 4: *Source separation results*

Dataset	SDR	PESQ	STOI
WSJ0-original	15.1	2.50	0.83
WSJ0-edges	15.2	2.61	0.82
WSJ0-edges-realistic	14.5	2.28	0.71

5. Discussion

Often the methodology for generating speaker positions in generated datasets is to make it completely random, but as discussed throughout this work, this is not realistic. Constraints such as enforcing a minimum distance between sources seem sensible at first but can yield unrealistic distributions. Without reporting either the separation distribution of the dataset or the performance of the source separation system with respect to the separation angle, it is difficult to compare results across different works, as we have shown the WER can change by over 73.5% relative by only changing the location of sources. We suggest that when generating simulated evaluation data to err towards sources being closer together rather than using a uniform distribution in order to more closely match real data. This work has focused on just one parameter of simulation design, however, other equally important parameters are often overlooked such as directivity patterns (i.e., the direction speakers are facing), the distance they are away from the microphone and the degree of speaker overlap [17].

6. Conclusions and future work

In this paper, we have employed automatic ways to estimate the angular distribution of speakers in a multi-speaker distant microphone scenario using face-detection and pose-detection techniques. We established confidence in these techniques by creating hand labels with the objective to evaluate the effectiveness of the tool on isolated frames and over the entire session. Using this analysis, we showed that in the CHiME-5 scenario where the camera has a field of view of 84.1 degrees, the speakers that are visible have an average angular separation of 17 degrees. We compared this distribution to common simulated datasets that are used to benchmark the state-of-the-art in speech enhancement and found there is a large disparity. We then showed that this disparity could have consequences for the research community, such as leading research down the wrong path by pursuing systems that optimise unrealistic angular separations.

In this work, we have analysed speaker position from the perspective of devices. This work is currently being extended to use the overlapping cameras in the CHiME-5 dataset to triangulate positions of speakers in the room, allowing the *distance* from the microphones to be estimated. Tracking techniques are also being explored to use the continuity between frames. Further to this, ways to integrate the video information into speech enhancement and ASR are being explored, but as highlighted in the analysis this is a challenging task as the modality often has missing information, e.g., people are not always facing the camera. The video data is not public, however, we are making the extracted detections and ground truth labels available[4].

[4] https://jackdeadman.github.io/chime5video

7. References

[1] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, J. Fischer, D. Hymes, S. T. Kimball *et al.*, "Proxemics," *Current anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[3] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2746–2760, 2019.

[4] R. Scheibler, E. Bezzam, and I. 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)*, 2018, pp. 351–355.

[5] J. Le Roux, S. Wisdom, H. Erdogan, and J. 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] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[7] Z.-Q. Wang, J. Le Roux, and J. 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.

[8] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 696–700.

[9] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech 2018*, 2018, pp. 1561–1565.

[10] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35.

[13] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

[14] N. Kanda, C. Böddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1248–1252.

[15] J. Du, Y.-H. Tu, L. Sun, L. Chai, X. Tang, M.-K. He, F. Ma, J. Pan, J.-Q. Gao, D. Liu, C.-H. Lee, and J.-D. Chen, "The USTC-NELSLIP Systems for CHiME-6 Challenge," in *CHiME-6 Workshop, Barcelona, Spain*, 2020.

[16] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6389–6393.

[17] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: dataset and analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2020.