INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Directional Speech Recognition for Speaker Disambiguation and Cross-talk Suppression

Ju Lin, Niko Moritz*, Ruiming Xie*, Kaustubh Kalgaonkar, Christian Fuegen, Frank Seide†*

Meta, USA

{julincs,nmoritz,seide}@meta.com

Abstract

With advances in mobile computing, smart glasses are becoming powerful enough to generate real-time closed captions of live conversations. Such system must distinguish speech from the conversation partner from the wearer's, and in public places it must not transcribe speech from unrelated bystanders to avoid confusion and to honor privacy.

We propose an end-to-end modeling approach that leverages the smart glasses' microphone array. But we go beyond beamforming for improved target-speaker SNR: We feed multiple audio channels *simultaneously* to a single ASR model as a basis for *speaker-attributed transcription* and *suppression of bystander cross-talk*.

Our proposed multi-channel directional ASR model processes multiple beamformer outputs for different steering directions *simultaneously* and combines it with serialized output training. Under room-acoustics and noise simulation, we demonstrate near perfect wearer/conversation-partner disambiguation and suppression of cross-talk speech from non-target directions.

Index Terms: speech recognition, multi-channel, multi-talker ASR, cross-talk suppression

1. Introduction

We introduce an ASR model that receives multiple audio input channels simultaneously. Besides SNR improvement, it leverages the multi-channel input more directly to disambiguate speakers from different directions and to suppress cross-talk. Automatically transcribing speech of a conversation partner at a distance of several feet is an important scenario—consider the automatic generation of captions for deaf or hard-of-hearing users. Background noise, reverberation, overlapping speech, and interfering speakers make this a challenging task. As a remedy, one can capture the speech with a *microphone array*—in a sense, that's what humans do. Microphone-array methods often aim to improve the SNR of target speech.

Literature roughly divides microphone-array based ASR into two categories: end-to-end approaches and hybrid approaches. In the end-to-end approaches [1, 2, 3, 4, 5], the multi-channel ASR model is optimized only via an ASR criterion with or without explicit separation modules. MIMO-speech [4] is a multichannel end-to-end neural network that defines source-specific time-frequency (T-F) masks as latent variables in the network, which in turn are used to transcribe the individual sources. In [5], MIMO-speech is further improved by incorporating an explicit localization sub-network. Recent studies [6, 7] in ASR and speaker separation have also investigated directly incorporating spatial features instead of using explicit sub-modules jointly trained with the ASR module. For instance, [7] proposes an "all-in-one" model where the 3D spatial feature is directly used as input to the ASR system without explicit separation modules.

Hybrid methods typically employ a pipeline-based paradigm, where a speech separation module explicitly separates the clean target speech or explicitly predicts speaker related masks [7, 8, 9, 10, 11]. For example, Chen et al. [8] proposed a method for estimating a target speaker mask with multi-aspect features that can extract the target speaker from a speech mixture. The extracted speech signal is then fed into ASR. However, such end-to-end and hybrid approaches for multi-channel ASR involve explicit speaker separation or masking, before inputs are fed input into the ASR system, or concatenating the spatial cues with the ASR features.

In contrast, our proposed approach utilizes *multiple superdirective beamformer outputs* for different steering directions, which are processed *simultaneously* by a *single* ASR encoder. This allows the system to *implicitly* perform speaker disambiguation and suppression (and some speech separation), by using directional information, effectively learning to compare the different beamformer outputs. Sometimes this is referred to as the cocktail-party effect [12]. One benefit is that this method does not use explicitly extracted speaker characteristics.

The microphone array in this work is a simulation of *Project Aria* smart glasses[13]. Project Aria glasses are prototype smart glasses with a broad range of sensors, available to research institutions and academia for research on ego-centric smart-glasses applications. Project Aria's microphone array consists of 7 channels as shown in Fig. 1. Although the work in this paper only used simulations instead of real Aria recordings, we tested the models with a real-time prototype. Anecdotally, it works reliably, consistent with the experimental results.

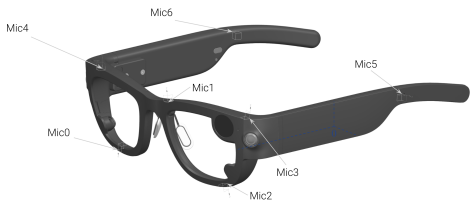

Figure 1: *Microphone locations on Project Aria glasses.*

2. Multi-channel Directional ASR

Fig. 2 shows the architecture of the proposed model. It consists of a front-end with multiple superdirective beamformers

*Equal contribution. †Corresponding author.

10.21437/Interspeech.2023-2076

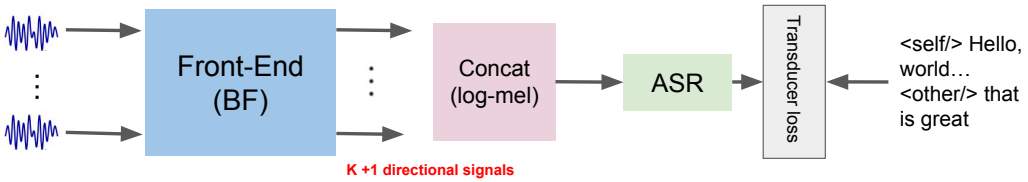

Figure 2: *Proposed directional speech recognition architecture.*

followed by a ASR module. The ASR module receives multiple input streams and is trained via serialized output training [14, 15] to detect speech from different directions. Unlike a standard single-channel ASR system, the multi-channel directional ASR system can leverage the differences in the directional outputs from the beamformers, allowing it to classify and separate speech signals arriving from different directions.

A straight-forward way is to feed all microphones' raw audio into N parallel front-ends, hoping that the model will automatically learn to separate speech from different directions. This, however, does not work: Our usual ASR feature extractors remove phase information—but temporal differences are the most important information for detecting direction of arrival. Instead, in this work, we pre-process the raw multi-channel audio by beamforming it for K horizontal steering directions around the smart-glasses device plus one in the speaker's mouth direction. These beamformers use predetermined coefficients. The ASR feature extraction front-end is then applied to these $K+1$ *beamformed* channels, the output of which are concatenated and fed into the ASR encoder neural network. This maps the problem from comparing phase differences to one of comparing magnitudes and feature characteristics derived from different steering directions.

As seen in Fig 2, the N channels of raw audio data are fed into the beamformer front-end, which then obtains the $K+1$ directional signals. We then extract the usual log-Mel features for each beamformer direction and concatenate them together. This concatenated vector constitutes the input of the ASR encoder.

The beamformers used in this work are *superdirective* beamformers [16, 17, 18]. A superdirective beamformer is derived by maximizing the directivity factor, or DF. Specifically, the method minimizes the power output and applies a linear constraint in order to obtain an undistorted output signal. This optimization maximizes the directivity index. The $K+1$ beamformers are predetermined. At runtime they are realized via one-dimensional convolutions.

Our ASR model is a Neural Transducer [19, 20, 21, 22]. This well-known end-to-end ASR architecture consists of three components: an encoder, a prediction network, and a joiner network. The goal of the transducer model is to produce a label sequence $Y = (y_1, \cdots, y_L)$ of length L, which can be a sequence of words or word-pieces, from an input sequence $X = (x_1, \cdots, x_S)$ of length S, typically a sequence of acoustic features like Mel-spectral energies. The encoder neural network processes the input sequence X and produces a sequence of acoustic representations, denoted as $H^{enc} = (h_1^{enc}, \cdots, h_T^{enc})$, of length T, which might differ from S due to sub-sampling. The prediction neural network acts as an internal language model or decoder to generate a representation h_u^{dec}, where u represents the decoder state. Generally, u depends on the previous output labels $y_{0:l-1} = (y_0, \cdots, y_{l-1})$, where y_0 corresponds to a start of sentence symbol, and l denotes the label index. Lastly, the joiner network takes the output representations from the encoder and prediction network as input and creates the joint representation $h_{t,u}^{joint}$, where t denotes the encoder frame index.

What sets our model apart is that we also incorporate *serialized output training*, or SOT [14, 15]. This is a technique for detecting speaker changes—in our case between the wearer and a target speaker (other)—as well as for recognizing partially overlapping speech. In our SOT implementation, ground-truth transcriptions from multiple speakers are sorted by the end times of all words. These are then interleaved, where at every speaker change, a special symbol ($\gg 0$ or $\gg 1$) is inserted. This way, the model learns to intersperse ASR transcripts with markup to indicate whether the speech came from *self* (the wearer of the glasses) or from *other* (the conversational partner opposite to the wearer). Note that compared to [14, 15], the availability of multiple input channels makes this task significantly easier.

Lastly, our model learns to suppress bystanders' speech: The ground truth for training the multi-channel ASR model includes only the transcripts of the *self* and *other* speakers. Speech of bystanders, that is, speech simulated from directions other than the target-speaker directions, is included in the training data as well, but with empty transcripts. This way, the model learns to ignore cross-talk. Experiments shows that this simple approach works with almost perfect accuracy.

Like [19], we used the alignment-restricted RNN-T (ARRNN-T) loss, which utilizes prior alignment information, such as forced alignment information from a traditional hybrid acoustic model, to limit the set of alignments to a valid subset. This results in significant improvements of memory usage and training speed.

3. Experiments and Results

3.1. Dataset

We conducted experiments using two datasets: the open-source Librispeech corpus [23], which consists of 960 hours of speech from audiobooks in the LibriVox project, and an in-house dataset of de-identified video data publicly shared by Facebook users. The training and evaluation sets of the in-house video data consist of 40k and 50 hours, respectively.

To simulate the training data, we generated 100,000 multi-channel room impulse responses (RIRs) for rooms with sizes ranging from [5, 5, 2] to [10, 10, 6] meters. We used the geometry of Aria glasses to simulate multi-channel data. Aria has 7 microphones. We generated the multi-channel signals using image-source methods (ISM)[24]. To better understand the impact of cross-talk on speech recognition, we generated four different training sets varying the locations of conversation partners and bystanders. Figure 3 provides details on the training settings. In the V1 configuration, the conversation partners are located between 1 and 11 o'clock (blue area), and bystanders are located between 1 and 11 o'clock (red area). The V2 and V3 settings leave a gap between the simulated partner and by-

stander directions. This is to study whether very close bystander and partner directions, such as in V1 and V4, might confuse the model during training due insufficient spatial resolution of the array.

We generated several test scenarios with different conditions for conversation partners and bystanders. There are three conversation-partner position sets: 12 o'clock, 11 or 1 o'clock, and 10 or 2 o'clock. We also varied the bystander placement, with four ranges: 3 to 5 o'clock, 7 to 9 o'clock, 2 and 10 o'clock, and 5 and 7 o'clock. This simulation resulted in 9 evaluation sets (3 partner conditions combined with 3 bystander conditions). Each contains 3367 utterances for Librispeech.

Noise from the public noise set was added to the clean audio segments in both the training and test sets, at SNRs ranging from -20 dB to 30 dB relative to the combined audio of wearer and partner, at intervals of 1 dB. The volume level of bystander speech, which varies with distance (e.g. roughly 30 dB attenuation at a distance of 4 feet), was randomly selected to be between approximately 6 and 36 dB, relative to the wearer. In addition, three overlap ratios between bystanders and main speakers are investigated: 0%, 50% and 100%. 0% indicates there is no overlap between bystander and main speakers.

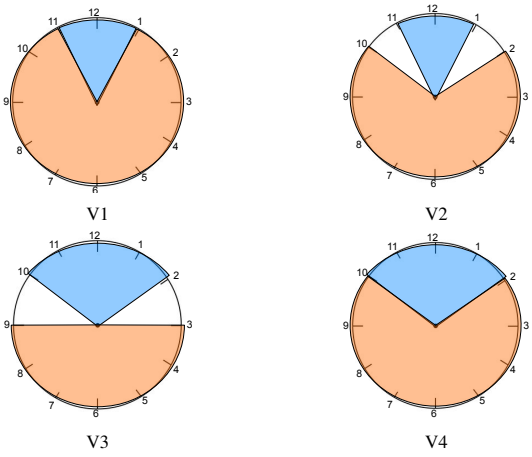

Figure 3: *Training configuration of spatial positions of the conversation partner and bystander. Blue areas represent the partner, while orange area represents the bystander area.*

3.2. Baseline systems

The baseline systems used for performance comparison are a single-channel ASR system and a interchannel phase differences (IPDs) system. The single-channel system takes the reference microphone signals (the first microphone) as input. IPDs [25, 6] can be calculated as follows

$$\mathrm{IPD}_{t,f}^{(n)} = \angle \mathrm{Y}_{t,f}^{n_1} - \angle \mathrm{Y}_{t,f}^{n_2}, \quad (1)$$

where $\angle \mathrm{Y}_{t,f}^{n_1}$ denotes the angle of the complex representation $\mathrm{Y}_{t,f}^{n_1}$ with t and f representing time and frequency and (n) is the index for the microphone pairs. The substation of the phase signals for a pair of microphone signal, including a target and a reference microphone, eliminates phase variations inherent in source signals and hence allows room acoustic characteristics to be directly captured. The IPD features are further augmented with magnitude spectra to leverage both spectral and spatial cues. For the short-term Fourier transform, we use a Hanning windows of 16 ms and a frame shift of 10 ms. 6 pairs of microphones are selected for IPD, which are (0,1), (0,2), (0,3), (0,4), (0,5), (0,6). The total dimension of the input feature after concatenation is 129 * (7+6) = 1677.

3.3. Model Setup

The baseline systems and the proposed systems are using the same model architecture, except for a different input dimension. For each beamformer direction or raw microphone channel, we extracted 80-dimensional log-Mel filterbank features. Input features from multiple directions or channels are concatenated. The *encoder network*'s input layer projects this resulting concatenated feature vector to 128 dimensions. Then, four consecutive frames are stacked to form a 512 dimensional vector (reducing the sequence length by 4x). This is followed by 20 Emformer blocks [26] with 8 attention heads and 2048-dimensional feed-forward layers. The RNN-T's *prediction network* contains three 512-dimensional LSTM layers with layer normalization and dropout. Lastly, the encoder and predictor outputs are both projected to 1024 dimensions and passed to an additive *joiner network*, which contains a linear layer with 4096 output BPE units.

We use an Adam optimizer with a tri-stage learning-rate scheduler. For LibriSpeech, models are trained for 120 epochs, with a base learning rate of 0.001, a warmup of 10,000 iterations, and forced annealing after 60 epochs. In additon, there is no external language model is used in our RNN-T model. For experiments on large-scale in-house data, a similar model architecture and training hyper-parameters were used, with training for 15 epochs.

3.4. Results

3.4.1. Beamformer analysis

Figure 4 depicts the beam patterns of the superdirective beamformer for 4 different directions, as indicated by the blue arrows. While beam patterns vary greatly, the gain is around 1 in the desired looking directions.

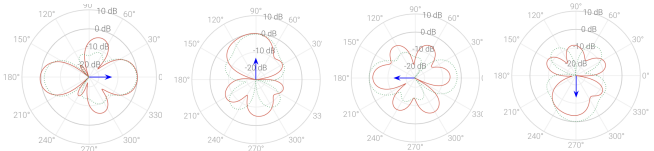

Figure 4: *Beam patterns at f = 2 kHz.*

3.4.2. Comparison with baseline systems

We will now compare our proposed directional ASR system (referred to as "D-ASR") with two baselines. The number of beams used in the ASR model,m $K+1$, was represented by the numbers in brackets after "D-ASR" - [D-ASR-1], [D-ASR-5], and [D-ASR-13]. These numbers denote the number of beams used in the ASR model, with "1" indicating beamformed output at 12 o'clock direction, "5" representing 4 beams for the horizontal plane (at 90-degree increments) plus the self-beam (to the wearer's mouth), and "13" representing 12 beams for the horizontal plane at 30-degree increments plus the self-beam.

Unless otherwise noted, word error rates (WERs) consider speaker attribution by counting *self* and *other* tags like words. A missing or incorrect speaker tag counts as one error.

First, Table 1 shows that the single beamformed input system (D-ASR-1) outperforms the single-channel reference-microphone system (SC-Raw mic) in most cases, by significant

Model	Partner [12]			Partner [11 / 1]			Partner [10 / 2]		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
D-ASR-5 (V1)	**12.0**	12.9	**11.9**	14.7	**15.2**	14.9	52.5	53.2	52.4
D-ASR-5 (V2)	**12.0**	13.8	12.0	14.7	16.1	14.7	52.7	53.9	52.6
D-ASR-5 (V3)	**12.0**	49.8	12.0	**14.2**	51.9	**13.9**	**14.5**	51.9	**14.5**
D-ASR-5 (V4)	**12.0**	36.8	12.0	14.2	38.5	14.1	15.4	43.8	15.2
IPD (V1)	15.0	15.7	14.7	15.9	16.4	16.1	53.8	54.3	53.8
IPD (V2)	14.9	16.8	14.7	15.2	16.8	15.3	53.7	54.8	53.7
IPD (V3)	15.1	55.5	14.9	15.3	54.6	15.1	15.6	54.9	15.7
D-ASR-1 (V1)	16.5	26.2	17.1	20.6	30.2	21.8	29.0	**39.5**	30.1
D-ASR-1 (V2)	17.2	29.0	16.3	19.7	31.5	20.6	28.3	41.4	28.7
D-ASR-1 (V3)	18.3	41.1	16.7	21.7	43.4	20.0	25.1	46.6	22.9
SC-Raw mic (V1)	41.3	42.1	40.9	40.9	41.6	42.1	41.3	41.1	40.9
SC-Raw mic (V2)	40.9	42.2	40.4	40.8	40.9	41.3	40.5	40.9	40.7
SC-Raw mic (V3)	41.5	43.0	40.8	42.0	42.3	42.4	41.8	41.9	41.3

Table 1: *WERs (%) for the proposed and baseline systems on Librispeech. **C1**: bystanders are located at 3 to 5 o'clock and 7 to 9 o'clock. **C2**: bystanders are located at 10 o'clock and 2 o'clock. **C3**: bystanders are located at 5 o'clock and 7 o'clock. Annotation is same to other tables. The overlap ratio is 0%.*

margins. I.e., using a single directional signal already provides valuable spatial cues. Compared with the strong IPD baseline systems, which uses explicit spatial cues, our proposed D-ASR with 5 directional signals consistently achieves better performance, demonstrating the effectiveness of our approach.

We also compared training conditions for different bystander locations. D-ASR is sensitive to unseen test conditions, similar to IPD-based methods. For example, the performance of D-ASR-5 with V3 training data drops significantly for the partner at 12 o'clock when bystanders are nearby the 10 and 2 o'clock directions (**C2**). In the V3 training data, bystanders are only located at 3 to 9 o'clock, so **C2** is an unseen condition. In contrast, the V4 training condition also includes bystanders close to the 10 and 2 o'clock directions, which results in improvements over V3, although the discrimination between bystander and partner is very narrow at this location.

3.4.3. Impact of the number of beams

We conducted ablation studies to measure the impact of the number of beams used for model training. Here, we fixed the model training using V4 configuration. Table 2 contains the results for Librispeech comparing four such systems. Comparing D-ASR-5 with D-ASR-1, we see that more beams can reduce the WER significantly on the conditions that bystanders are far away from the partners (**C1** and **C3**). When bystanders are close to the partner (C2), D-ASR-1 performs somewhat better, likely because the spacial resolution of the beamforming is not sufficient to resolve bystander and partner directions that are very close. Similar to using 13 beams, we also see an improvement in the **C1** and **C3** conditions. Applying volume perturbation further boosts ASR performance, which should teach the model to not rely on amplitude differences to discriminate speaker directions but on other special and spectral cues instead.

3.4.4. Impact of overlap ratio; speaker attribution accuracy

Next, we investigated the impact of performance under different overlap conditions. As presented in Table 3, we initially validated the performance of our D-ASR model under ideal

Model	Partner [1/11]		
	C1	C2	C3
D-ASR-1 (V4)	21.1	38.5	19.7
D-ASR-5 (V4)	14.2	38.5	14.1
D-ASR-13 (V4)	13.4	41.8	13.5
D-ASR-13 (V4) + vol. Perturb	**13.3**	**36.7**	**13.2**

Table 2: *The impact of the different number beams for the directional speech recognition. The overlap ratio is 0%.*

Model	Over-lap	Noise	Cross-talk	Partner [1/11]	
				C1	C3
SC-Raw (V1)	0%	Y	Y	40.9	42.1
D-ASR-13 (V4)	0%	N	N	5.5	5.5
		Y	N	12.2	12.5
		Y	Y	12.8	13.0
D-ASR-13 (V4)	50%	N	N	5.5	5.5
		Y	N	13.2	12.9
		Y	Y	14.2	14.0
D-ASR-13 (V4)	100%	N	N	5.6	5.6
		Y	N	14.1	14.3
		Y	Y	16.0	15.9

Table 3: *WER (%) at varying ratios of overlap of cross-talk with self/other speech.*

conditions, i.e., no noise and cross-talk, in which it achieved around 5.5% on Librispeech test-clean dataset in all cases. At 0% overlap, the cross-talk speech increases the total amount of speech by approximately 50%—undesired speech that should not be recognized (with cross-talk disabled, audio length still increases by 50%, but of silence or noise). The single-channel model only suppresses some lower-volume cross-talk, while it decodes its majority as insertion errors, pushing the WERs to over 40%. Whereas the D-ASR model suppresses cross-talk almost perfectly at the lower overlap ratios: At 0% overlap, the WER increases from 12.2 to 12.8%, 0.6% absolute, corresponds to only about 1.2% of the cross-talk audio! Accuracy degrades a bit more at 100% overlap, when bystander speech effectively becomes background noise.

Let's use the D-ASR-13 (v4) 0% no cross-talk C3 configuration to look at speaker-attribution accuracy. We split ASR output/ground truth by speaker tags. Now, words attributed to the wrong speaker become insertions or deletions. After this split, the resulting WER increases from 12.5% to 12.7%. Hence, speaker attribution works almost perfectly as well.

3.4.5. Results on large-scale dataset

Finally, we conducted experiments on our large-scale in-house dataset. As shown in Table 4, we observed similar tendency on the in-house data. The proposed directional ASR model consistently outperforms the IPD baseline system in all cases.

Model	Partner [12]			Partner [11 / 1]		
	C1	C2	C3	C1	C2	C3
D-ASR-5 (V1)	11.0	11.0	11.0	12.8	13.1	12.5
D-ASR-5 (V2)	10.8	11.3	10.8	11.1	11.5	11.1
D-ASR-5 (V3)	11.1	56.7	11.1	11.1	59.9	11.1
IPD (V1)	-	-	-	22.2	23.0	22.2
IPD (V2)	-	-	-	22.3	23.6	22.2
IPD (V3)	-	-	-	22.2	69.5	22.1

Table 4: *WER (%) on our in-house data, at overlap ratio 0%.*

4. Conclusions

This paper has introduced an ASR modeling approach that uses multi-channel directional input. Besides the usual SNR improvement, multi-channel audio is leveraged more directly to disambiguate speakers from different directions and to suppress cross-talk. Our RNN-T based model is trained to annotate speaker changes and ignore bystander speech in an end-to-end fashion. Comprehensive experiments were conducted for conversational ASR with smart glasses using different bystander and conversational partner conditions. We have demonstrated that the proposed directional ASR system disambiguates the wearer's from the conversation partner's speech and suppresses bystander speech (from undesired directions) almost perfectly.

5. References

[1] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5325–5329.

[2] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.

[3] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *International conference on machine learning*. PMLR, 2017, pp. 2632–2641.

[4] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-speech: End-to-end multi-channel multi-speaker speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 237–244.

[5] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, Y. Xu, S.-X. Zhang, and D. Yu, "Directional ASR: A new paradigm for e2e multi-speaker speech recognition with source localization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8433–8437.

[6] 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.

[7] Y. Shao, S.-X. Zhang, and D. Yu, "Multi-channel multi-speaker ASR using 3D spatial feature," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6067–6071.

[8] 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 *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[9] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4819–4823.

[10] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

[11] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5739–5743.

[12] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[13] "Project aria," https://about.meta.com/realitylabs/projectaria/.

[14] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming multi-talker ASR with token-level serialized output training," *arXiv preprint arXiv:2202.00842*, 2022.

[15] X. Chang, N. Moritz, T. Hori, S. Watanabe, and J. L. Roux, "Extended graph temporal classification for multi-speaker end-to-end ASR," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7322–7326.

[16] G. W. Elko, S. Gay, and J. Benesty, "Superdirectional microphone arrays," *KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE*, pp. 181–238, 2000.

[17] G. Huang, J. Benesty, and J. Chen, "Superdirective beamforming based on the krylov matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2531–2543, 2016.

[18] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007.

[19] J. Mahadeokar, Y. Shangguan, D. Le, G. Keren, H. Su, T. Le, C.-F. Yeh, C. Fuegen, and M. L. Seltzer, "Alignment restricted streaming recurrent neural network transducer," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 52–59.

[20] N. Moritz, F. Seide, D. Le, J. Mahadeokar, and C. Fuegen, "An investigation of monotonic transducers for large-scale automatic speech recognition," *arXiv preprint arXiv:2204.08858*, 2022.

[21] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6059–6063.

[22] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao *et al.*, "Developing rnn-t models surpassing high-performance hybrid models with customization capability," *arXiv preprint arXiv:2007.15188*, 2020.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210.

[24] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[25] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: Spectrogram vs waveform separation," *arXiv preprint arXiv:1905.07497*, 2019.

[26] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6783–6787.