



# Directional Speech Recognition with Full-Duplex Capability

Ju Lin, Yiteng Huang, Ming Sun, Frank Seide, Florian Metzger

<sup>1</sup>Meta, USA

{julincs, yah}@meta.com

## Abstract

Recent work on directional automatic speech recognition (DASR) has enabled automatic transcription of a conversation partner several feet away via smart glasses. DASR leverages multiple microphones in the glasses by using multiple beamformers simultaneously. We aim to make the DASR insensitive to scenarios that also involve text-to-speech (TTS) playback. This could enable additional future scenarios like simultaneous speech translation.

How to prevent ASR from capturing the system's own TTS output, while maintaining optimal clarity of the captured conversation partner's speech? We experiment with two modern linear acoustic echo cancellation (AEC) algorithms. To remedy accuracy regressions from echo residuals, we propose *AEC-aware model training*.

While AEC alone eliminates most TTS loopback, dramatically improving the word-error rate (WER) by over 70%, AEC-aware model training provides further relative WER boosts of 13% or more.

**Index Terms:** directional speech recognition, full-duplex, multi-microphone, acoustic echo cancellation

## 1. Introduction

Wearable devices, such as smart glasses, are equipped with amazing computational capability to seamlessly transcribe speech, e.g. to generate real-time closed captions for live conversations. This capability is of great importance in the domain of Automatic Speech Recognition (ASR), particularly offering valuable benefits for hearing-impaired users. Prior research [1, 2, 3, 4] has demonstrated that Directional Automatic Speech Recognition (DASR), by leveraging a multi-microphone array embedded in smart glasses, proves effective in discerning speakers among the wearer, the conversation partner, and unrelated bystanders. Furthermore, DASR also demonstrates superior robustness to noise when compared to ASR systems utilizing single-channel beamformed signals.

In this study, we expand the scope of DASR by enabling glasses wearers to receive system output in the form of text-to-speech (TTS). The TTS output, e.g., simultaneous speech translation, can be fully overlapped with conversation partner's speech, which forms a *full-duplex* scenario. A key challenge is to prevent the ASR system from capturing the system's own TTS output, while maintaining optimal clarity in capturing the conversation partner's speech. What adds on to this challenge is that the device loudspeakers may be located in very close proximity to some of the device's microphones, while the conversation partner is usually situated several feet away. This results in very poor signal-to-echo ratios (SER) between the conversation

partner and the device playback. This is commonly addressed by Acoustic Echo Cancellation, or AEC.

In the present context, the DASR framework posits specific constraints on the AEC module. Specifically, the AEC operations must be linear to prevent interference with phase information. AEC precedes beamforming, which is crucial for capturing directional representations in the DASR system. Hence, we examine the effect of two linear, multi-channel AEC algorithms. To address the residual echo issue accompanied with linear AEC operations, we further introduce the *AEC-aware model training* strategy, where AEC-processed data is integrated into the training stage using a multi-style training approach. Our experiments conducted on real-world collected test data demonstrate the effectiveness of both AEC approaches and AEC-aware model training in enhancing the performance of multi-channel DASR systems.

Related work on AEC for speech recognition includes [5], which proposed an implicit acoustic echo cancellation (iAEC) framework where a neural network is trained to exploit the additional information from a reference microphone channel to learn to ignore the interfering signal and improve detection performance. In [6], a Conformer-based waveform-domain neural AEC model is explored, in which the model is trained by jointly optimizing Negative Scale-Invariant SNR (SISNR) and ASR losses on a large speech dataset. A neural multi-channel AEC was proposed in [7] to achieve echo removal for all microphones by leveraging one single deep neural network (DNN) model.

Our contributions are summarized as follows:

- We investigate two simple yet effective linear AEC algorithms for far-field speech recognition. To the best of our knowledge, our work is one of the earliest efforts exploring AECs on smart glasses for multi-channel ASR.
- We conduct a comprehensive analysis to quantify the impact of AEC on the multi-talker ASR task, providing valuable insights into the relationship between AEC and ASR performance.
- We propose AEC-aware model training for multi-talker ASR systems, demonstrating the effectiveness of our proposed approach and highlighting the potential benefits of integrating AEC into ASR model training.

## 2. Directional ASR with AEC

Fig. 1 illustrates the system architecture of our DASR system with AEC. It is comprised of multi-channel AEC module, concerted beamformers introduced in [1], and a streaming RNN-T [8, 9, 10, 11] based ASR system trained with serialized output training, or SOT [12, 13]. We will describe AEC component in

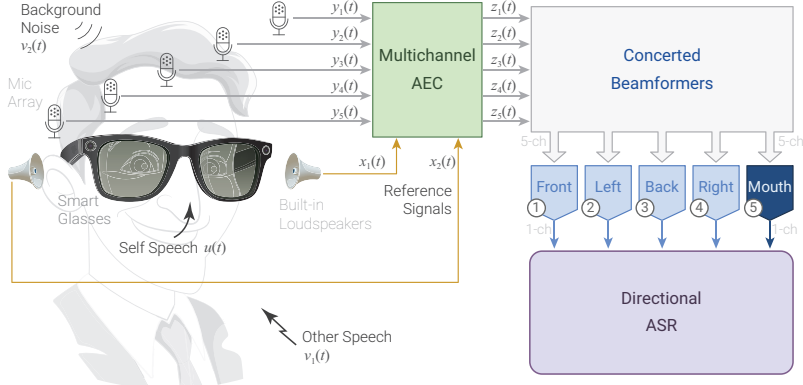


Figure 1: System diagram of directional automatic speech recognition with acoustic echo cancellation.

detail in the following subsection.

## 2.1. AEC Approaches

Acoustic echo cancellation (AEC) tackles a fundamental system identification challenge, particularly focusing on the so-called loudspeaker-enclosure-microphone (LEM) configuration [14]. This system is typically represented by a single time-varying linear filter, with the adaptive finite impulse response (FIR) model being a prevalent choice in practical applications.

Let  $h_p$  ( $p = 1, 2, \dots, P$ ) denote the channel impulse response from the  $p$ th out of  $P$  loudspeakers to a microphone. Then the microphone output is expressed as

$$y(t) = \sum_{p=1}^P h_p * x_p(t) + s(t), \quad (1)$$

where  $t$  is the discrete time index,  $*$  indicates linear convolution,  $x_p(t)$  is the  $p$ th reference signal, and  $s(t)$  is the mixture of user speech  $u(t)$ , other speech  $v_1(t)$ , and background noise  $v_2(t)$ . If  $h_p$  has only  $N$  taps, then (1) can be written in a more concise vector-matrix form:

$$y(t) = \sum_{p=1}^P \mathbf{h}_p^T \mathbf{x}_p(t) + s(t) = \mathbf{h}^T \mathbf{x}(t) + s(t), \quad (2)$$

where

$$\begin{aligned} \mathbf{h}_p &\triangleq [h_{p,0} \ h_{p,1} \ \dots \ h_{p,N-1}]^T, \\ \mathbf{x}_p(t) &\triangleq [x_p(t) \ x_p(t-1) \ \dots \ x_p(t-N+1)]^T, \\ \mathbf{h} &\triangleq [\mathbf{h}_1^T \ \mathbf{h}_2^T \ \dots \ \mathbf{h}_P^T]^T, \\ \mathbf{x}(t) &\triangleq [\mathbf{x}_1^T(t) \ \mathbf{x}_2^T(t) \ \dots \ \mathbf{x}_P^T(t)]^T, \end{aligned}$$

and  $(\cdot)^T$  denotes the transpose of a vector or a matrix.

There exists a plethora of adaptive algorithms designed for AEC, with the normalized least-mean-square (NLMS) [15, 16] being the most widely known for its simplicity and numerical stability. Using the NLMS, we update the estimate of  $\mathbf{h}$  at time  $t$  with

$$\hat{\mathbf{h}}(t) = \hat{\mathbf{h}}(t-1) + \frac{\mu}{\mathbf{x}^T(t)\mathbf{x}(t) + \delta} \mathbf{x}(t)e(t), \quad (3)$$

where  $\mu$  is the step size that controls the rate of adaptation,  $\delta$  is a small positive constant to avoid division by zero, and

$$e(t) \triangleq y(t) - \hat{\mathbf{h}}^T(t-1)\mathbf{x}(t)$$

is the prior error signal.

In practical AEC applications, the NLMS algorithm faces a notable obstacle: its convergence tends to be quite sluggish, particularly when dealing with lengthy impulse responses and colored inputs such as speech. An effective technique to deal with long impulse responses is frequency-domain adaptation [17]. Using the frequency-domain NLMS algorithm [18] allows leveraging the fast Fourier transform (FFT) to minimize computational cost. This is the approach adopted in our in-house Dual-Path-AEC implementation. This AEC, referred to as such, derives its name from its utilization of the two-path algorithm [19, 20], in which a background filter adapts as in a conventional echo canceler and a foreground canceler does the actual cancellation. The coefficients of these two filters in each frequency band can be copied in either direction whenever a decision logic declares that the background canceler is performing better than the current foreground canceler.

It is widely acknowledged that subband echo cancellation is another technique commonly employed to deal with long impulse responses [21, 22]. Rather than employing a multi-rate filter bank for signal decomposition and re-synthesis [23], we opt to utilize the short-time Fourier transform (STFT) method, primarily due to its widespread availability in various libraries.

Using a  $K$ -point STFT analysis, a linear convolution in (1) is rigorously converted into a sum of  $K$  cross-band filter convolutions in the STFT domain, which are necessary to cancel the aliasing caused by downsampling in each frequency subband [24]. This produces When  $K$  is large, the so-called multiplicative transfer function (MTF) approximation [25].

But since we want to represent the long impulse responses with short analysis windows (smaller  $K$ ), the convolutive transfer function (CTF) approximation [26] is more accurate and less restrictive:

$$\begin{aligned} Y(k, n) &= \sum_{p=1}^P \sum_{l=0}^{L-1} H_p(k, l) X_p(k, n-l) + S(k, n) \\ &= \mathbf{h}^T(k) \mathbf{x}(k, n) + S(k, n), \end{aligned} \quad (4)$$

where

$$\begin{aligned} \mathbf{h}(k) &\triangleq [\mathbf{h}_1^T(k) \ \mathbf{h}_2^T(k) \ \dots \ \mathbf{h}_P^T(k)]^T, \\ \mathbf{x}(k, n) &\triangleq [\mathbf{x}_1^T(k, n) \ \mathbf{x}_2^T(k, n) \ \dots \ \mathbf{x}_P^T(k, n)]^T, \\ \mathbf{h}_p(k) &\triangleq [H_p(k, 0) \ H_p(k, 1) \ \dots \ H_p(k, L-1)]^T, \\ \mathbf{x}_p(k, n) &\triangleq [X_p(k, n) \ X_p(k, n-1) \ \dots \ X_p(k, n-L+1)]^T. \end{aligned}$$

To explore the performance ceiling, we prioritized quality over computational efficiency and employed the most ambitious recursive least squares (RLS) algorithm to solve (4) for an esti-

mate of  $\mathbf{h}(k)$  in each frame:

$$\hat{\mathbf{h}}(k, n) = \mathbf{R}_{xx}^{-1}(k, n)\mathbf{r}_{xy}(k, n), \quad (5)$$

where

$$\begin{aligned} \mathbf{R}_{xx}(k, n) &= \lambda \mathbf{R}_{xx}(k, n-1) + (1-\lambda)\mathbf{x}(k, n)\mathbf{x}^H(k, n), \\ \mathbf{r}_{xy}(k, n) &= \lambda \mathbf{r}_{xy}(k, n-1) + (1-\lambda)\mathbf{x}(k, n)Y^*(k, n), \end{aligned}$$

are the approximations (using exponentially weighted moving average with a forgetting factor  $0 < \lambda < 1$ ) of  $E\{\mathbf{x}(k, n)\mathbf{x}^H(k, n)\}$  and  $E\{\mathbf{x}(k, n)Y^*(k, n)\}$ , respectively. Here  $(\cdot)^*$  denotes the conjugate of a complex variable,  $(\cdot)^H$  denotes the Hermitian transpose of a vector or matrix, and  $E\{\cdot\}$  denotes the mathematical expectation.

In our research, we refer to this design as STFT-RLS AEC. The forgetting factor is determined by

$$\lambda = e^{-\frac{1}{\tau \cdot f_s}}, \quad (6)$$

where  $\tau$  is the RLS's time constant and  $f_s$  is the STFT's frame rate.

## 2.2. AEC-Aware Model Training

Speech data after AEC will inevitably introduce residual echoes which might cause accuracy degradation in the ASR phase given ASR models have not encountered such residuals during training. To address this potential issue, we introduce an AEC-aware multi-channel ASR training approach. This approach involves generating simulated training data with TTS playback, followed by processing it with an AEC algorithm, such as Dual-Path-AEC, to minimize the echo, in the same way as would be done at inference time. Subsequently, we fine-tune a pre-trained model, initially trained solely on original multi-channel data, using an augmented corpus, including composition of 15% AEC-processed data and 85% normal multi-channel data. Through this approach, the ASR model adapts to recognize speech in the presence of residual echo effects introduced by the AEC algorithms.

# 3. Experiments and Results

## 3.1. Dataset

Models are trained on an in-house dataset of 14.6k hours of video data with single-channel audio. As real-world multi-channel training data of sufficient amounts is not available, all multi-channel training data is simulated. We used the geometry of 5-microphone prototype glasses similar to the Aria glasses [27] as in [1]. We first generate 1M multi-channel room impulse responses (RIRs) using image-source methods (ISM)[28] via the "pyroomacoustics" library [29], with room sizes ranging from [5, 5, 2] to [10, 10, 6] meters. Subsequently, we simulate training data by situating single-channel audio clips in space, representing the wearer ("self"), the conversation partner ("other"), and unrelated bystanders. This simulation mirrors a conversation between the self and the other, introducing some overlap, and simulating bystander crosstalk. For detailed information on the simulation process, refer to [1].

To simulate TTS playback in the training data, we utilize the measured loudspeaker-to-microphone impulse response (IR) to simulate how the TTS output feeds back into the microphones. Subsequently, we add the left-channel and right-channel simulated playback signals to generate a 5-channel audio representation. The "TTS signal" is randomly selected from

an English ASR dataset. We then mix the speech and playback signals at a random SER ranging from -10 to 10 dB. It's worth noting that the SER pertains to the back microphone signals, which are the closest to the loudspeaker. The final output of the simulation comprises 5-channel mixed signals, along with the original 1-channel input signal, serving as the loopback signal for AEC processing.

In addition, noise from the DNS Challenge [30] was added to the clean audio segments in training, at SNRs ranging from -5 to 30 dB w.r.t. the combined audio of wearer and partner, at intervals of 1 dB.

To evaluate the ASR performance in presence of TTS + AEC, we collected real-world data which consists of conversations between a wearer wearing the prototyping glasses and a conversation partner at a distance of around 4 to 6 feet. We then mix the playback signals with speech signals at various SER levels representing different playback volumes.

## 3.2. Model setup

The model configuration closely follows [2]. For each beamformer direction, 80-dimensional log-Mel filterbank features are extracted. These features from all channels (steering directions) are then fed into the Convolutional front-end, which comprises 2 conv2d blocks, each with 5 channels, filters of size  $2 \times 5$ , and a stride setting of  $1 \times 2$ . Subsequently, six consecutive frames are stacked, forming a 320-dimensional vector, effectively reducing the sequence length by a factor of 6x. This is succeeded by 20 Emformer layers [31], each incorporating 4 attention heads and 2048-dimensional feed-forward layers.

The RNN-T prediction network includes a single 256-dimensional LSTM layer with layer normalization and dropout. Finally, both the encoder and predictor outputs are projected to 768 dimensions and passed through an additive joiner network, which consists of a ReLU followed by a linear layer with 9001 output Sentence-Piece-based units. All models undergo training for 8 epochs, utilizing an Adam<sub>sam</sub> optimizer, a tri-stage learning-rate scheduler with a base learning rate of 0.0005, and a warm-up of 10,000 batches. For AEC-aware model training, a pre-trained AEC-unaware model is fine-tuned with an additional epoch.

## 3.3. Evaluation Metrics

The AEC methods are evaluated using the commonly used *perceptual evaluation of speech quality* (PESQ) score [32, 33, 34], the *short-time objective intelligibility* (STOI) score [35]. The performance of AEC approaches are also evaluated in terms of utterance-level echo return loss enhancement (ERLE) [36].

Unlike common ASR tasks, the ASR system used here also has the task of identifying who is speaking (SELF = the wearer or OTHER = conversation partner), plus it must not transcribe speech from unrelated bystanders. This is reflected by a modified word-error rate (WER) metric, which, in addition to Substitution, Insertion, and Deletion errors, also counts words attributed to the wrong speaker as an error. This is denoted by SA for "speaker attribution" error. Words from unrelated bystanders accidentally transcribed count as Insertion errors, and correct words incorrectly attributed to bystanders count as Deletion errors.

## 3.4. Results

### 3.4.1. Performance evaluation of AEC algorithms

First, we conducted ablation studies to assess the influence of various parameters in the STFT-RLS AEC system. The results,

Table 1: Speaker attribution (SA) error and attributed ("self", "other") word error rates (WER) on simulated test data for three SER conditions.

Model	-5dB (SER)				0dB (SER)				5dB (SER)				Clean (no TTS)			
	OTHER		SELF		OTHER		SELF		OTHER		SELF		OTHER		SELF	
	WER	SA	WER	SA	WER	SA	WER	SA	WER	SA	WER	SA	WER	SA	WER	SA
AEC-unaware																
w/o AEC	66.3	49.3	51.4	0.3	63.4	47.5	42.1	0.3	60.6	44.0	36.6	0.1	10.4	0.6	9.3	0.4
Dual-Path-AEC	18.0	1.2	22.4	2.5	15.3	1.1	21.6	1.8	14.1	0.6	21.4	1.6	-	-	-	-
STFT-RLS-AEC	<b>17.3</b>	0.5	<b>10.2</b>	0.3	<b>13.3</b>	0.5	<b>9.6</b>	0.2	<b>12.2</b>	1.0	<b>9.7</b>	0.2	-	-	-	-
AEC-aware																
Dual-Path-AEC	<b>13.4</b>	1.6	<b>9.1</b>	0.3	11.8	0.6	9.0	0.3	10.8	0.9	9.0	0.3	10.0	0.9	9.2	0.4
STFT-RLS-AEC	13.8	0.4	9.4	0.4	<b>11.5</b>	0.7	9.0	0.4	<b>10.6</b>	0.6	9.0	0.4	11.4	0.9	9.0	0.4

Table 2: WER (%) comparison of different configurations for the STFT-RLS AEC method.

RLS Time Const $\tau$ (s)	Frame Size $K$	Filter Length $L$	OTHER		SELF	
			WER	SA	WER	SA
3	512	4	19.0	1.1	10.1	0.3
3	1024	4	<b>17.3</b>	0.5	10.2	0.3
3	2048	8	17.7	1.0	10.9	0.4
3	3072	8	18.7	0.6	11.3	0.5
4	1024	4	18.0	0.7	<b>10.0</b>	0.3
4	1024	8	17.8	0.6	10.2	0.3
4	2048	8	17.7	1.1	10.7	0.4
4	3072	8	17.4	0.7	10.6	0.4

Table 3: Performance comparison of Dual-Path and STFT-RLS AECs on ERLE (in dB), PESQ, and STOI for 5 channels.

SER (dB)	Ch. No.	Dual-Path AEC			STFT-RLS AEC		
		ERLE	PESQ	STOI	ERLE	PESQ	STOI
-5	1	3.77	2.92	0.51	3.37	2.95	0.86
	2	2.88	3.15	0.49	2.26	3.07	0.94
	3	3.06	3.09	0.49	2.26	3.09	0.93
	4	6.97	2.24	0.45	6.20	2.22	0.87
	5	7.01	2.18	0.45	6.30	2.23	0.86
5	1	2.46	3.69	0.52	2.41	3.89	0.97
	2	2.04	3.74	0.50	1.98	4.10	0.98
	3	2.12	3.74	0.50	2.04	4.09	0.98
	4	4.17	3.10	0.48	3.94	3.41	0.94
	5	4.22	3.05	0.48	3.98	3.39	0.94

outlined in Table 2, reveals critical influence each of the parameter has on the model performance; for instance, given  $\tau = 3$  seconds and  $L = 4$ , increasing the Frame Size  $K$  from 512 to 1024 improves WER for Other significantly (1.7% absolute). Subsequently, in our ensuing experiments involving STFT-RLS AEC, we opted for a RLS time constant of  $\tau = 3$  seconds, a frame size of  $K = 1024$ , and a filter length of  $L = 4$ , striking a balance between accuracy and responsiveness.

Next, we conducted a comparative analysis of the performance between the Dual-Path and STFT-RLS AEC methods, focusing on perceptual quality and intelligibility. The results, presented in Table 3, indicate that Dual-Path-AEC exhibits superior perceptual quality as measured by ERLE. Conversely, STFT-RLS-AEC surpasses Dual-Path-AEC in terms of STOI, which is indicative of better intelligibility. Previous research suggests that STOI is more closely associated with ASR tasks, a correlation reaffirmed by our subsequent ASR experiments. In terms of PESQ, we observe that STFT-RLS AEC achieves comparable performance with Dual-Path AEC at the low SER condition (-5dB) but is superior than Dual-Path AEC at the high SER condition (5dB).

### 3.4.2. Performance comparison of AEC for DASR

First, we assessed the impact of incorporating AEC-processed data on model accuracy for clean data without TTS echo, as depicted in Table 1. the Dual-Path-AEC model not only sustains its performance but also exhibits a slight improvement of approximately 0.4%, particularly in the "OTHER" category, with an additional epoch of fine-tuning. Conversely, for STFT-RLS AEC, we observed a 1% degradation in the "OTHER" category.

Subsequently, we evaluated both AEC approaches across various signal-to-echo ratios. Both methods effectively mitigated echo across all levels, as demonstrated in Table 1, compared to conditions without AEC. However, an AEC-convergence issue surfaced during the initial processing seconds for Dual-Path-AEC, leading to the misclassification of TTS playbacks as SELF, resulting in significantly higher WER compared to STFT-RLS AEC (22.4% vs. 10.2% under -5dB SER condition).

In our final examination, we tested the AEC-aware training strategy, revealing its effectiveness for both AEC approaches, as illustrated in Table 1. This strategy notably reduced the WER for the "OTHER" category from 18.0% to 13.4% for Dual-Path and from 17.3% to 13.8% for STFT-RLS under the most challenging conditions. Additionally, AEC-aware training partially addressed the convergence issue in the Dual-Path-AEC algorithm which can be inferred from the significant improvement of self speaker WER. Moreover, the AEC-aware models showcased impressive performance for the self-speaker, even under a -5dB condition, with WERs of 9.1% for Dual-Path AEC and 9.4% for STFT-RLS AEC, compared to 9.3% in the clean (without TTS) condition, respectively.

## 4. Conclusion

This paper addresses a practical challenge in directional Automatic Speech Recognition (ASR) when confronted with Text-to-Speech (TTS) playbacks on smart glasses. We proposed the integration of a multi-channel Acoustic Echo Cancellation (AEC) into Directional ASR (DASR). Two AEC approaches were explored, both demonstrating effectiveness in eliminating TTS echoes. This enhancement resulted in a substantial improvement in the baseline Word Error Rate (WER) for the "OTHER" speaker, achieving a relative reduction of around 70% (up to 80%) compared to the configuration without AEC. Incorporating an AEC-aware model training strategy further improved the model performance.

Acknowledgment: The authors would like to thank Vahid Khanagha, Kaustubh Kalgaonkar, Yang Liu, and Hoang Do for the valuable discussions on AEC approaches.

## 5. References

- [1] J. Lin, N. Moritz, Y. Huang, R. Xie, M. Sun, C. Fuegen, and F. Seide, "Agadir: Towards array-geometry agnostic directional speech recognition," *arXiv preprint arXiv:2401.10411*, 2024.
- [2] J. Lin, N. Moritz, R. Xie, K. Kalgaonkar, C. Fuegen, and F. Seide, "Directional speech recognition for speaker disambiguation and cross-talk suppression," in *Proc. INTERSPEECH*, 2023, pp. 3522–3526.
- [3] 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.
- [4] T. Feng, J. Lin, Y. Huang, W. He, K. Kalgaonkar, N. Moritz, L. Wan, X. Lei, M. Sun, and F. Seide, "Directional source separation for robust speech recognition on smart glasses," *arXiv preprint arXiv:2309.10993*, 2023.
- [5] S. Cornell, T. Balestri, and T. Sénéchal, "Implicit acoustic echo cancellation for keyword spotting and device-directed speech detection," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1052–1058.
- [6] S. Panchapagesan, A. Narayanan, T. Z. Shabestary, S. Shao, N. Howard, A. Park, J. Walker, and A. Gruenstein, "A conformer-based waveform-domain neural acoustic echo canceller optimized for asr accuracy," *arXiv preprint arXiv:2205.03481*, 2022.
- [7] H. Zhang and D. Wang, "Multi-channel and multi-microphone acoustic echo cancellation using a deep learning based approach," *arXiv preprint arXiv:2103.02552*, 2021.
- [8] J. Mahadeokar, Y. Shanguan, 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.
- [9] N. Moritz, F. Seide, D. Le, J. Mahadeokar, and C. Fuegen, "An investigation of monotonc transducers for large-scale automatic speech recognition," *arXiv preprint arXiv:2204.08858*, 2022.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] M. M. Sondhi, "Adaptive echo cancelation for voice signals," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer, 2007, ch. 45, pp. 903–927.
- [15] —, "An adaptive echo canceler," *Bell Syst. Tech. J.*, no. 46, pp. 497–511, 1967.
- [16] B. Widrow and M. E. Hoff Jr., "Adaptive switching circuits," in *IRE Wescon Conf. Rec.*, New York, 1960, pp. 96–104.
- [17] M. Dentino, J. McCool, and B. Widrow, "Adaptive filtering in the frequency domain," *Proc. IEEE*, no. 66, pp. 1658–1659, 1978.
- [18] E. R. Ferrara Jr., "Fast implementation of LMS adaptive filter," *IEEE Trans. Acoust. Speech Signal Process.*, no. 28, pp. 474–475, 1980.
- [19] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE Trans. Commun.*, no. Com-25(6), pp. 589–595, 1977.
- [20] E. J. Diethorn, "Improved decision logic for two-path echo cancelers," in *Proc. IEEE Workshop on Acoustic Echo and Noise Control*, New York, 2001.
- [21] I. Furukawa, "A design of canceller of broad band acoustic echo," in *Proc. Int. Teleconf. Symp.*, New York, 1984, pp. 232–239.
- [22] W. Kellermann, "Kompensation akustischer echos in frequenzteilbandern," in *Proc. Aachener Kolloquium*, L. Butzer, Ed., RWTH Aachen, 1984, pp. 322–325.
- [23] —, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE ICASSP*, 1998, pp. 2570–2573.
- [24] M. Portnof, "Time frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Process.*, vol. ASSP-28, pp. 55–69, Feb. 1980.
- [25] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, pp. 337–340, May 2007.
- [26] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 546–555, May 2008.
- [27] K. Somasundaram, J. Dong, H. Tang, J. Straub, M. Yan, M. Goelese, J. J. Engel, R. De Nardi, and R. Newcombe, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.
- [28] 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.
- [29] 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)*. IEEE, 2018, pp. 351–355.
- [30] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matussevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *INTER\_SPEECH*, 2020.
- [31] 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.
- [32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Proc.*, Salt Lake City, UT, May 2001, pp. 749–752.
- [33] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-P recommendation P.862, Feb. 2001.
- [34] *Perceptual objective listening quality prediction*, ITU-P recommendation P.863, Mar. 2018.
- [35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [36] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic press library in signal processing*. Elsevier, 2014, vol. 4, pp. 807–877.