INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

On Training a Neural Residual Acoustic Echo Suppressor for Improved ASR

Sankaran Panchapagesan, Turaj Zakizadeh Shabestary, Arun Narayanan

Google LLC, U.S.A.

{panchi,turajs,arunnt}@google.com

Abstract

Acoustic Echo Cancellation (AEC) is critical for accurate recognition of speech directed at a smart device playing audio. Previous work has showed that neural AEC models can significantly improve Automatic Speech Recognition (ASR) accuracy. In this paper, we train a conformer-based waveform-domain neural model to perform residual acoustic echo suppression (RAES) on the output of a linear AEC. We focus specifically on improving ASR accuracy in realistic mismatched test conditions, when training on large-scale simulated training data, as needed for production voice-interaction systems. Our key finding is that instead of naively using the best evaluation-time linear AEC configuration during neural RAES model training, using a weaker linear AEC generalizes significantly better, with 17-30% lower word error rate (WER) on a realistic re-recorded test set. Overall, the neural RAES model yields 38-53% WER reduction over the linear AEC alone.

Index Terms: Acoustic Echo Cancellation, Waveform Neural AEC, Residual Echo Suppression, TasNet, ASR

1. Introduction

Acoustic Echo Cancellation (AEC) is an essential algorithm used to enhance the speech input to smart speakers that are playing audio such as text-to-speech (TTS) responses, audiobooks or music. The digital assistant typically performs keyword spotting or ASR only on the enhanced speech output by the AEC. Conventional AEC systems typically estimate a linear filter between the playback reference signal and the received echo at the microphone [1], and are quite effective at improving the signal-to-interference ratio of the speech input.

Residual echo in the linear AEC output can significantly hinder speech and hotword recognition, particularly at higher playback volumes and when the playback audio is predominantly speech (e.g., TTS, podcasts). Residual echo could be due to non-linearities in the device loudspeaker, long room reverberation times, or the speaker moving. Several neural network based methods have recently been proposed for AEC and residual echo suppression, with architectures based on LSTMs, dilated convolutions, attention networks, U-Nets and others [2–14]. However, almost all of these approaches are focused on optimizing and improving speech enhancement and quality metrics, which are not matched with our goal of ASR.

Recently there has been more work focused on neural AEC for speech or hotword recognition [15–17]. In [15], an LSTM-based neural AEC model was proposed, taking microphone and reference logmel features as input and predicting enhanced logmel features. A novel auxiliary *ASR loss*, defined as the L2 loss between ASR encoder output representations computed from target and predicted logmel features, was shown to reduce ASR WER by 12-17%. In [16], a very similar ASR loss was used to improve WER by 9-11% with slight degradation in speech quality, for a neural model operating as a residual echo suppressor in a fullband AEC system. In [17], temporal convolution networks with input mixture and reference features, and predicting posteriors for keyword spotting or device directed speech detection, were shown to be more parameter efficient than predicting enhanced features. However this implicit AEC approach is narrowly focused on and tied to the tasks of interest.

In [18], a waveform-domain neural AEC model optimized for ASR was presented, with an architecture inspired by the TasNet model [19], and using conformer layers [20] for the enhancement mask estimation. One advantage of a waveform-domain model is that the same model can potentially be used for different applications like ASR and hotword by including corresponding auxiliary loss functions. The model in [18] was trained on a large speech dataset simulated with both synthetic and real echoes as interference, and the auxiliary ASR loss from [15] was shown to be effective in improving WER. By cascading the waveform-domain neural AEC model after a linear adaptive AEC system, significant WER reductions of 56-59% over the linear AEC alone were demonstrated on a realistic re-recorded test set, although the systems were not trained jointly.

Motivated by the work in [18], the current work proposes neural residual acoustic echo suppression (RAES) in the waveform domain, by training the model on the output of the linear AEC instead of the microphone signal, keeping the playback reference as an auxiliary input. Unlike conventional approaches that train the RAES model using linear AEC settings that are found to be optimal in test conditions, we find that using a weaker linear AEC system during training is key for better generalization and performance on mismatched test sets. On a realistic re-recorded test set, evaluating with a state-of-the-art ASR model that includes an endpointer optimized for latency [21], the proposed training for neural RAES models using a weaker linear AEC gives 17-30% word error rate (WER) reduction over using a strong linear AEC during training, and 38-53% WER reduction over the linear AEC alone.

The rest of the paper is organized as follows. Section 2 briefly discusses related work. In Section 3, the architecture and training loss of the proposed waveform-domain neural RAES model are described. In Section 4, the experimental setup including training and evaluation data are described. Experimental results and ablations are presented in Section 5, and conclusions and ideas for future work in Section 6.

2. Related Work

In [6], a Conv-TasNet model trained on input reference and linear AEC output was proposed for residual echo suppression. In [8], EchoFilter, a TasNet-style masking neural AEC model

10.21437/Interspeech.2023-2209

using attention and LSTM modules was proposed, that included an auxiliary double-talk detection network. In [5], a neural RAES model consisting of a contextual attention module between recurrent encoder and decoder layers was proposed, taking log-spectra of mixture, reference and output error signal of the linear AEC as inputs. As mentioned above, these neural models were optimized only for enhancement and speech quality, unlike our model which targets ASR accuracy.

The ICASSP 2022 AEC challenge [22] included ASR accuracy as an evaluation metric and has stimulated relevant research. In [23], a gated convolutional F-T-LSTM neural network post-filter was trained after a linear AEC, and trained with an echo-aware loss and using double-talk detection as an auxiliary task. In NeuralEcho [24], a fully neural 2-stage AEC system using attention based recurrent neural networks was developed for joint echo and noise suppression, also potentially incorporating speaker-aware enhancement and AGC. In [25], a 2-stage neural model was proposed, consisting of a multi-channel AEC and a joint AEC-beamformer performing double-talk detection. Both [24] and [25] used 2nd order statistics such as cross-correlation between microphone and reference signals as input. In [26], a conformer-based front-end operating on logmel features was developed to jointly handle AEC, multi-channel enhancement and ID-based speaker separation.

This paper, to the best of our knowledge, is the first to optimize a conformer-based waveform-domain RAES model for ASR accuracy, and to address the important issue of generalization to mismatched test data by analyzing how the training time configuration of linear AEC affects evaluation time ASR quality.

3. Waveform-domain Neural RAES Model

3.1. Model Architecture

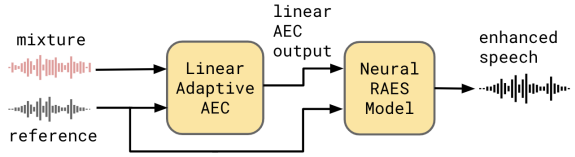

Figure 1: *Cascade of Linear Adaptive AEC and Neural Residual Acoustic Echo Suppressor (RAES).*

Figure 1 shows the block diagram of our system, with the neural RAES model cascaded after a linear adaptive AEC system. The neural model takes the linear AEC output signal concatenated with the playback reference as input and outputs enhanced speech that is intended for downstream applications such as hotword and ASR.

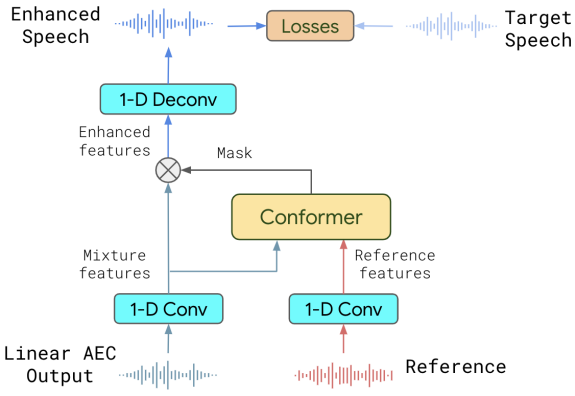

Figure 2: *Waveform-domain neural RAES architecture.*

Figure 2 shows our proposed architecture for the waveform-domain neural RAES model, which is the same as that of the neural AEC in [18], with a simplified figure for clarity. The model performs enhancement by masking learned features, and is inspired by TasNet [19]. The input mixture and reference signals are converted to features by separate 1-D convolution layers. The mixture (here the linear AEC output) and reference features are stacked and input into a conformer-based mask estimator network which estimates an enhancement mask in the feature domain. The mask is multiplied with the mixture features to produce predicted features, which are converted into the predicted waveform by the 1-D deconvolution layer. Conformer layers combine convolutional and transformer layers to efficiently model both global and local dependencies in audio signals. Further details on the conformer layers can be found in [18, 20, 21, 27].

3.2. Loss Functions

The model is trained with a combination of Negative SISNR loss and the ASR loss proposed in [15]. The front-end that computes logmel features needed for the ASR loss is implemented as a non-trainable layer that backpropagates gradients from the ASR loss layer during model training. The total training loss is:

$$\mathcal{L} = -\text{SISNR}(\mathbf{s}, \hat{\mathbf{s}}) + \lambda \mathcal{L}_{\text{ASR}} \tag{1}$$

where λ is a hyperparameter. SISNR is defined as the SNR obtained after scaling the target signal to have least squares error with the predicted signal. The SISNR formula and derivation may be found in [19,28]. $\mathcal{L}_{\text{ASR}}$ is defined as an L2 loss between target and predicted ASR encoder output sequences:

$$\mathcal{L}_{\text{ASR}} = \sum_k \left\| \mathbf{E}_{\text{ASR}}(\mathbf{S}_k) - \mathbf{E}_{\text{ASR}}(\hat{\mathbf{S}}_k) \right\|_2^2 \tag{2}$$

where $\mathbf{S}_k$ and $\hat{\mathbf{S}}_k$ are feature vectors at frame k computed by the front-end from the target and predicted signals respectively, and $\mathbf{E}_{\text{ASR}}(\cdot)$ is the ASR encoder function. For the experiments in this paper, we used the conformer encoder from the ASR model described in [21].

4. Experiments

4.1. Training Data

The neural RAES models is trained on target utterances selected from a 50.8k hours dataset consisting of Librispeech [29] (960 hours), LibriVox[1] (46.5k hours), and internal vendor collected datasets (3.3k hours). As recommended in [15,18], target speech is mixed with both synthetic and real echoes to create mixture signals at signal-to-echo ratio (SER) between -20 dB and 5 dB. Real echoes are obtained by rerecording utterances from Librispeech and internal text-to-speech (TTS) model training data, on smart speakers from multiple rooms. The second dataset is chosen since an important use case is to cancel TTS responses played by the device. For synthetic echoes, utterances from Librispeech and Librivox, and noise snippets from Getty[2] and YouTube Audio Library[3] are convolved with synthetic room impulse responses (RIRs), with close microphone location to mimic device playback. Target speech is also convolved with RIRs to simulate farfield conditions.

4.2. Simulated Librispeech Test Set

We use simulated Librispeech test sets created from the Librispeech test-clean set by artificially adding reverberation and noise to the target speech utterances, and mixing in held-out rerecorded echoes at SERs of 5dB, 0dB, -5dB and -10dB. Each

[1] https://librivox.org
[2] https://www.gettyimages.com/about-music
[3] https://youtube.com/audiolibrary

set contains 2620 utterances with around 52.5k words. Similar to the training utterances, the associated playback reference is stored with each test utterance.

4.3. Re-recorded Test Set

We also use a more realistic test set created by re-recording speech queries played out from a second loudspeaker towards the device under playback, and also simultaneously capturing the playback reference. The speech queries were recorded from three different speaker distances of 1.3m, 3.3m and 5.2m, and device playback volume setting varied over a range up to the maximum setting of 10. Each test set utterance has a total length up to 30 sec, beginning with around 10 sec of echoed reference only, followed by the target speech query mixed with interfering echo, and ending with several seconds of continuing echo only. The test set was partitioned based on playback volume and speaker position into Easy, Moderate and Difficult sets, containing respectively around 19k, 12.3k, and 11.4k utterances, and around 141k, 92k, and 85k words. This test set is very challenging in general due to high interfering echo levels, and particularly for the neural AEC/RAES models due to significant mismatch with the simulated model training data.

4.4. Linear Adaptive AEC

Our Linear AEC system performs subband adaptive filtering using STFT similar to [30], but uses longer STFT frames (128ms) and within-band only filters of order 4. These values were chosen based on latency constraints for ASR, and on ASR and hotword accuracy on development sets.

Table 1: *Linear AEC Parameters, their evaluation-time stronger values, and weaker values used during Neural RAES training (see Table 3).*

Parameter	Eval value	Weaker value
FIR Order	4	1
Filter Update Interval	1.5 sec	3 sec
Frame Overlap	75%	50%
Forgetting Factor	0.995	0.98
Alignment Threshold	0.2	0.1
Max Ref-Mic Alignment Lag	550 ms	60 ms
Max Mic Alignment Buffer Length	2000 ms	500 ms
Max Ref Alignment Buffer Length	2000 ms	500 ms

4.5. Weaker linear AEC parameters during model training

As will be reported in Section 5.2, we found that using the same linear AEC configuration parameters tuned for best test set performance also while training the neural RAES model did not yield sufficient improvements on the re-recorded test set. We found that a strategy of making the linear AEC significantly weaker during training helped the model generalize much better to the re-recorded test set. The weaker training settings included a predictor filter order of 1 instead of 4, less frequent filter updates, and shorter alignment search windows between microphone and reference signals. Table 1 shows the list of linear AEC parameters and their weaker values during training.

Our hypothesis for the effectiveness of this approach is that since the re-recorded test set is mismatched with the model training data, using the strong linear AEC during training likely results in overfitting. The weaker linear AEC system also results in larger levels of residual echo, and echo not cancelled for some portion of the training data due to lack of alignment between microphone and reference signals. This allows the model to handle residual echo better when they occur in unseen test conditions, as is also shown in Section 5.2.

4.6. Waveform-domain Neural RAES

The input signals to the waveform-domain neural RAES model are framed by the 1-D convolution layers using windows of length 5ms (80 samples at 16kHz sampling rate), shifted by 2.5ms (40 samples). The learned feature dimension is 128, which is also the dimension of the 4 conformer layers of the mask estimator. In our experience, the learned feature dimension needs to be larger than the window length to get good results with the TasNet architecture. The convolutional blocks in the conformer layers use a kernel size of 15, while the causal attention has 8 heads with a left context of 31 frames. The 1-D deconvolution layer uses tanh activation to produce audio samples in the range $(-1, 1)$. The total size of the model is 1.6M parameters. Given the 31 frames of left-context of the attention in the four conformer layers, and the frame shift of 2.5ms, the waveform-domain neural RAES model uses a total past context of approximately $4 \times 31 \times 2.5 = 310$ms.

We used the Lingvo toolkit [31] to train models. For the neural RAES model, the ASR loss weight (λ in Equation 1) was fixed to be 1e3 for all experiments below. While larger values of λ do give some WER gains (not reported here), the focus of this paper is on the choice of linear AEC parameters during RAES model training, which yields much larger WER gains. During training, the ASR loss weight is increased linearly, starting from 0.0 at 20k steps to the selected value at 100k steps, and kept fixed after that.

4.7. ASR Evaluations

We use two different ASR models for evaluations. The first ASR model is an LSTM RNN-T model [32] trained on $\sim$400k hours of English speech from domains like VoiceSearch, YouTube, Telephony and Farfield. The 512-dimensional model input features are obtained by stacking four successive frames of 128-dimensional logmel features computed from 32 msec windows with 10 msec hop, and then subsampling by a factor of 3. Training utterances were anonymized and hand-transcribed, and the model also uses data augmentations like SpecAug [33] and simulated noise [34]. For inference with the first ASR model, we use label-synchronous beam search, with no endpointer. The second ASR model is an improved state-of-the-art conformer-based transducer model [21] trained on the same datasets and feature front-end as the LSTM RNN-T model above, and is approximately the same size, with $\sim$115M parameters. The conformer encoder has twelve 512-dimensional layers with masked self-attention using 23 left-context frames. The model decoding parameters were optimized for both accuracy and latency, and endpointer decisions are made using both an acoustic voice activity detector and an end-to-end end-of-speech prediction model [35]. Note that the ASR models are not jointly trained with the AEC model, and are kept frozen during training and inference.

5. Results

5.1. Results on simulated Librispeech test sets

We first evaluated the different AEC systems on the simulated Librispeech test sets described in Section 4.2 using the LSTM RNN-T ASR model described in Section 4.7. This matched test set was used mainly to sanity-check the implementation.

Table 2: *WERs(%) with different AEC/RAES models on simulated Librispeech test subsets at varying SERs.*

AEC Method	5 dB	0 dB	-5 dB	-10 dB
Linear Adaptive AEC	19.1	24.9	26.8	28.8
Waveform Neural AEC [18]	**9.8**	**11.5**	**15.3**	21.9
Linear + Waveform Neural RAES	13.6	14.4	15.5	**17.8**

The results of the different AEC / RAES methods on the simulated Librispeech test sets are presented in Table 2. The ASR loss weight for the neural AEC model was λ=5e4. It is seen that both the neural AEC and the linear AEC + neural RAES cascade perform significantly better than the linear AEC at all SERs. These test sets are more challenging for the linear AEC, since the utterances do not begin with a echo-only segment where the adaptive filter can converge before the mixed speech starts. The neural AEC seems to work well on matched data at higher SERs, while the neural RAES model performs better in lower SER conditions, wherein the residual echo levels are higher. As mentioned above, these results on matched test data mainly serve to provide a sanity check.

5.2. Results on Re-recorded Test Set with Cascaded System

We next evaluated the proposed waveform-domain neural RAES model on the more realistic re-recorded test set described in Sec. 4.3. As mentioned there, unlike the simulated Librispeech test sets above, this test set is significantly mismatched with the simulated model training data, but includes an echo-only segment before the target query for the linear AEC to converge. Hence, while the neural AEC alone performs poorly (WERs $> 100\%$) on this test set, cascading after the linear AEC was very effective. On the other hand, the neural RAES model has been specifically trained to be cascaded with the linear AEC for precisely this use case of mismatched realistic test data. The evaluations here used the state-of-the-art Conformer-based ASR model with endpointer described in Section 4.7.

Table 3: *WERs with ASR model with endpointer on re-recorded test sets, with linear adaptive AEC only, vs. cascade of linear AEC and neural AEC [18] and RAES (this paper) models.*

AEC Method	Easy	Moderate	Difficult
Linear AEC only	14.5	31.0	63.6
Linear + Waveform Neural AEC [18]	10.8	21.5	50.8
Linear + Waveform Neural RAES:			
Trained w/ Strong Linear AEC	10.0	20.7	50.0
Trained w/ Weak Linear AEC	**8.3**	**14.5**	**39.1**

WER results on the re-recorded test set are presented in Table 3. The challenging nature of the test set is clear from the poor performance of the linear AEC alone. It is seen that the cascade of linear AEC and the previously proposed independently trained waveform-domain neural AEC model [18] gives large reductions in WER over the linear AEC alone. In our first attempt at jointly training a linear AEC + waveform neural RAES cascade, the strong evaluation-time values (see Table 1) were used for the linear AEC parameters during training. The results are shown in the third row, where it is seen that joint training yields only marginal improvements over [18]. From the next row of the table, we see that when we instead use a weak linear AEC system (see Table 1) during training, the linear AEC + waveform neural RAES cascade performs significantly better than using a strong linear AEC during training, with WER reductions of 17%, 30% and 22% on the Easy, Moderate and Difficult test set partitions, respectively.

Overall the linear AEC + waveform neural RAES cascade gives large ASR accuracy improvements over the linear AEC alone, with WER reductions of 43%, 53% and 38% on the Easy, Moderate and Difficult test sets, respectively. Note that the strong linear AEC system is still used during all evaluations. The results of Sections 5.1 and 5.2 also show that the neural RAES model generalizes well to different ASR models, a useful property since the RAES model and the ASR model can be independently optimized and maintained in a production setting.

5.3. Ablation experiments with Linear AEC Parameters

We next performed some ablation experiments by taking the RAES model trained with weak linear AEC in Table 3, and studying the effect of changing one linear AEC parameter at a time. For this experiment, since the microphone-reference alignment lag and buffer lengths are co-dependent, they were not varied. The WER results are shown in Table 4. Note that during evaluation only strong linear AEC parameters were used.

Table 4: *Ablation experiments with Linear AEC Parameters. Effect on WERs on re-recorded test sets, by changing one linear AEC parameter at a time, from a weak value to a strong value during training (see Table 1).*

Linear + Waveform Neural RAES	Easy	Moderate	Difficult
Trained w/ Weak Linear AEC	8.3	14.5	39.1
Change:			
FIR Order to 4	10.6	18.5	45.5
Alignment Threshold to 0.2	7.7	14.7	40.0
Forgetting Factor to 0.995	7.9	**14.0**	38.8
Frame Overlap to 75%	7.7	14.2	37.8
Filter Update Interval to 1.5 sec	**7.5**	14.2	**37.2**
Trained w/ Strong Linear AEC	10.0	20.7	50.0

As can be seen from the second row of Table 4, changing the FIR filter order from 1 to 4 has the most negative impact on WER, and may explain most of the WER gap between weak and strong linear AEC training parameters. From Table 4 we see that strengthening the other parameters one at a time seems to result in comparable performance, and sometimes even improvement. For example, with all weak linear AEC parameters except the filter update interval of 1.5 sec, we obtain 2-10% lower WERs than the all-weak configuration, 25-31% lower WERs than the all-strong configuration, and 41-54% lower WERs than the linear AEC alone. Changing the microphone-reference alignment lag and buffer length values in a tied manner seems likely to yield worse results than the all-weak configuration, and needs to be studied in future work.

6. Conclusions

In this paper, the problem of training a waveform-domain neural RAES model for improved ASR was studied, with a particular focus on generalization to mismatched test sets. The model architecture used conformer layers in a Tasnet-style masking approach, and was trained by jointly optimizing Negative SISNR and ASR losses on a large speech dataset simulated with both synthetic and real echoes as interference. A key finding was that using a weaker instead of a strong linear AEC system during the training of the neural RAES model yields significantly better ASR performance on mismatched test sets. On a realistic re-recorded test set, significant reductions in WER of 17-30% were demonstrated using this training approach. The neural RAES model gives 38-53% WER reduction over the linear AEC alone. Future work would include allowing linear AEC parameters during training to vary randomly over specified distributions, incorporating double talk detection into the neural model, and optimizing the model also for hotword accuracy.

7. Acknowledgements

We wish to thank James Walker, Nathan Howard, Alex Park, Tom O'Malley and Joe Caroselli for discussions.

8. References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Wiley-Interscience, 2004.

[2] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Interspeech*, 2018.

[3] Q. Lei, H. Chen, J. Hou, L. Chen, and L. Dai, "Deep Neural Network Based Regression Approach for Acoustic Echo Cancellation," in *ICMSSP*, 2019.

[4] A. Fazel, M. El-Khamy, and J. Lee, "Deep Multitask Acoustic Echo Cancellation," in *Interspeech*, 2019.

[5] ——, "CAD-AEC: Context-Aware Deep Acoustic Echo Cancellation," in *ICASSP*, 04 2020, p. 6919–6923.

[6] H. Chen, T. Xiang, K. Chen, and J. Lu, "Nonlinear residual echo suppression based on multi-stream conv-tasnet," *arXiv preprint arXiv:2005.07631*, 2020.

[7] J.-H. Kim and J.-H. Chang, "Attention wave-u-net for acoustic echo cancellation." in *Interspeech*, 2020, pp. 3969–3973.

[8] L. Ma, S. Yang, Y. Gong, X. Wang, and Z. Wu, "Echofilter: End-to-end neural network for acoustic echo cancellation," *arXiv preprint arXiv:2105.14666*, 2021.

[9] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *ICASSP*, 2021.

[10] L. Ma, S. Yang, Y. Gong, and Z. Wu, "Multi-scale attention neural network for acoustic echo cancellation," *arXiv preprint arXiv:2106.00010*, 2021.

[11] H. Chen, G. Chen, K. Chen, and J. Lu, "Nonlinear residual echo suppression based on dual-stream dprnn," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–11, 2021.

[12] K. N. Watcharasupat, T. N. T. Nguyen, W.-S. Gan, S. Zhao, and B. Ma, "End-to-end complex-valued multidilated convolutional neural network for joint acoustic echo cancellation and noise suppression," *arXiv preprint arXiv:2110.00745*, 2021.

[13] H. Zhang and D. Wang, "Neural cascade architecture for multi-channel acoustic echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2326–2336, 2022.

[14] X. Zhou and Y. Leng, "Residual acoustic echo suppression based on efficient multi-task convolutional neural network," *arXiv preprint arXiv:2009.13931*, 2020.

[15] N. Howard, A. Park, T. Z. Shabestary, A. Gruenstein, and R. Prabhavalkar, "A neural acoustic echo canceller optimized using an automatic speech recognizer and large scale synthetic data," in *ICASSP*, 2021.

[16] H. Zhao, N. Li *et al.*, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *IEEE ICASSP*, 2022.

[17] S. Cornell, T. Balestri, and T. Sénéchal, "Implicit acoustic echo cancellation for keyword spotting and device-directed speech detection," *arXiv preprint arXiv:2111.10639*, 2021.

[18] S. Panchapagesan, A. Narayanan, T. Z. Shabestary *et al.*, "A conformer-based waveform-domain neural acoustic echo canceller optimized for asr accuracy," in *Interspeech*, 2022.

[19] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[21] B. Li, A. Gulati, J. Yu, T. N. Sainath, C.-C. Chiu, A. Narayanan, S.-Y. Chang, R. Pang, Y. He, J. Qin *et al.*, "A better and faster end-to-end model for streaming asr," in *ICASSP*, 2021.

[22] R. Cutler, A. Saabas, T. Parnamaa *et al.*, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP*, 2022.

[23] S. Zhang, Z. Wang, J. Sun *et al.*, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP*, 2022.

[24] M. Yu, Y. Xu, C. Zhang, S. Zhang, and D. Yu, "Neuralecho: A self-attentive recurrent neural network for unified acoustic echo suppression and speech enhancement," *arXiv preprint arXiv:2205.10401*, 2022.

[25] V. Kothapally, Y. Xu, M. Yu, S. Zhang, and D. Yu, "Joint neural aec and beamforming with double-talk detection," *Interspeech*, 2022.

[26] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A conformer-based asr frontend for joint acoustic echo cancellation, speech enhancement and speech separation," in *IEEE ASRU*, 2021.

[27] A. Narayanan, C.-C. Chiu, T. O'Malley, Q. Wang, and Y. He, "Cross-attention conformer for context modeling in speech enhancement for asr," *arXiv preprint arXiv:2111.00127*, 2021.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP*, 2019.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[30] Y. Avargel and I. Cohen, "Performance analysis of cross-band adaptation for subband acoustic echo cancellation," in *International Workshop on Acoustic Echo and Noise Control*, 2006.

[31] J. Shen, P. Nguyen, Y. Wu, Z. Chen *et al.*, "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," *arXiv preprint arXiv:1902.08295*, 2019.

[32] T. N. Sainath, Y. He *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP*, 2020.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[34] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Interspeech*, 2017.

[35] B. Li, S.-Y. Chang, T. N. Sainath *et al.*, "Towards fast and accurate streaming end-to-end asr," in *ICASSP*, 2020.