SCA: STREAMING CROSS-ATTENTION ALIGNMENT FOR ECHO CANCELLATION

Yang Liu, Yangyang Shi, Yun Li, Kaustubh Kalgaonkar, Sriram Srinivasan, Xin Lei

Meta, US
{yangliuai, yyshi, yunli1, kaustubhk, sriramsri, leixin}@meta.com

ABSTRACT

End-to-End deep learning has shown promising results for speech enhancement tasks, such as noise suppression, dereverberation, and speech separation. However, most state-of-the-art methods for echo cancellation are either classical DSP-based or hybrid DSP-ML algorithms. Components such as the delay estimator and adaptive linear filter are based on traditional signal processing concepts, and deep learning algorithms typically only serve to replace the non-linear residual echo suppressor. This paper introduces an end-to-end echo cancellation network with a streaming cross-attention alignment (SCA). Our proposed method can handle unaligned inputs without requiring external alignment and generate high-quality speech without echoes. At the same time, the end-to-end algorithm simplifies the current echo cancellation pipeline for time-variant echo path cases. We test our proposed method on the ICASSP2022 and Interspeech2021 Microsoft deep echo cancellation challenge evaluation dataset, where our method outperforms some of the other hybrid and end-to-end methods.

Index Terms— echo cancellation, delay estimation, complex attention

1. INTRODUCTION

While the use of voice communication has seen rapid growth, cancelling acoustic echo without suppressing the near-end talker remains a major unsolved problem in providing high quality speech. Traditionally, digital signal processing (DSP) based linear echo cancellation has been applied based on estimating the acoustic echo path with an adaptive filter. This approach fails when the echo path is time-varying or non-linear which, results in either echo leaks or significant suppression of the near-end talker during double-talk.

In recent years, deep neural network (DNN) based acoustic echo cancellation (AEC) methods have achieved a significant improvement over the traditional signal processing based methods. Deep complex convolution recurrent network (DC-CRN) designed for noise suppression [1] could be modified for the AEC task to better learn the relationship between frequency bands for effectively suppressing echo [2]. The main drawback of inplace DC-CRN is the larger number of parameters. Recently, Indenbom et al. propose a self-attention alignment for AEC, which is capable of handling non-aligned microphone and far-end signals in linear and non-linear echo path scenarios [3]. In most cases, researchers assume that the echo path is linear and the time delay is limited to a known prior and effectively combine traditional signal processing with a neural network. Wang et al. use a deep feed-forward sequential memory network (DFSMN) as a post-filter after an adaptive filter based linear AEC [4]. However, the performance of the existing AEC algorithms, especially those with low complexity, may be greatly degraded in real-life practical applications [5, 6]. In these applications, software-related latency or hardware-related latency may lead to large time-variant delays.

Inspired by the traditional signal processing alignment method such as cross-correlation [7] and Emformer [8], we propose an end-to-end real-time streaming deep neural network without any extra alignment module. Compared to prior works, our proposed streaming cross-attention alignment is used at the beginning of CRN network to improve the baseline model behaviour. The SCA-CRN has three contributions. First, cross-attention is applied to use near-end microphone signal to align the far-end signal. This work adds multi-head cross attention together with other components in the transformer [9] like layer norm, feedforward neural (FFN) layer and projection operations. Second, real and imaginary information are considered as two independent tensors to avoid complex computation, but the real and imaginary tensors share the weight of attention layers. Compared to complex-value attention used in [3], SCA is straightforward to implement in real-time applications. Third, SCA is implemented with a streaming mask to limit the cross-attention to access very limited look-ahead context in training which supports the models for low-latency streaming applications.

2. PROPOSED METHOD

2.1. Problem formulation

For a generic AEC system, we define the microphone signal as $d(n)$ which consists of near-end speech $s(n)$, acoustic echo $z(n)$ and background noise $v(n)$:

$$d(n) = s(n) + z(n) + v(n), \tag{1}$$

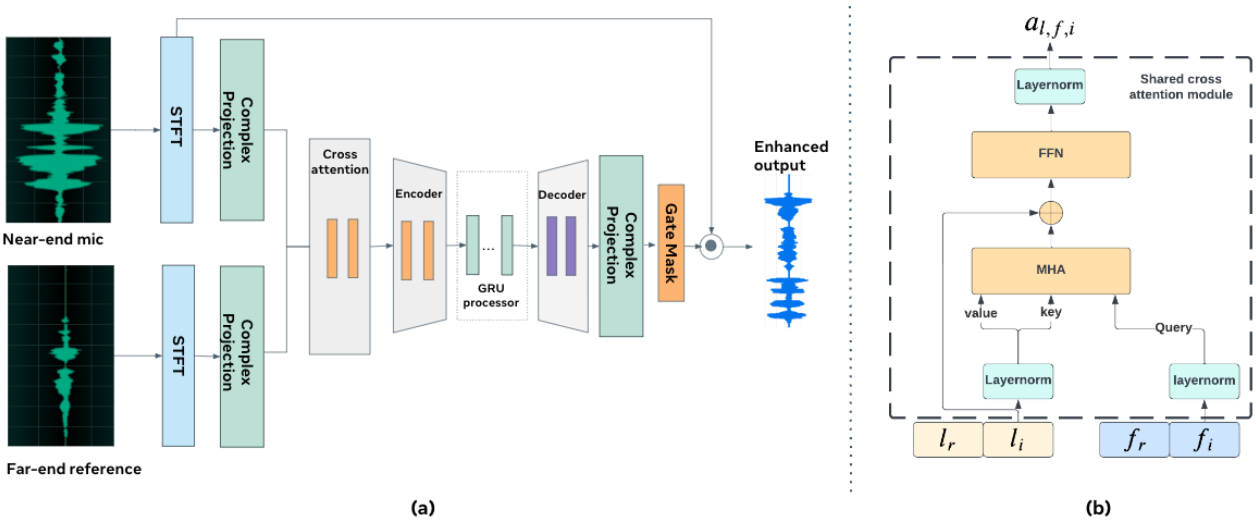

Fig. 1. (a) streaming cross-attention alignment AEC network pipeline. (b) The calculation of the streaming cross-attention attention $a_{l,f,i}$. The cross attention from the imaginary $a_{l,f,i}$ and the real $a_{l,f,r}$ share the same cross-attention module. The cross attention from the far-end and near-end $a_{f,l,i}$ and $a_{f,l,r}$ uses another cross-attention module with same architecture.

where n refers to the time sample index. $z(n)$ is a delayed version of the far-end reference signal $x(n)$ via echo path with potential nonlinear distortions caused by loudspeakers. This delay is related to the echo propagation path between the microphones and loudspeakers, hardware and software. This means the delay is time varied, unknown and difficult to estimate. The AEC task aims to separate $s(n)$ apart from $d(n)$, on the premise that "unaligned" $x(n)$ is known. The error signal $e(n)$ and linear echo $y(n)$ are generated using $x(n)$ and $d(n)$ by using standard adaptive filtering techniques.

2.2. Network architecture

The network structure is modified based on DC-CRN [1]. The SCA-CRN is composed by the SCA module, encoder, recurrent attention module and decoder, as shown in Figure 1(a). All audio signals are sampled at 48kHz.

The network consists of two branches to be input into the SCA module. They are complex projection of the near-end microphone recording and the far-end reference recording. For running on edge devices, the models do not use complex tensors. The real and imaginary components are considered as two independent real tensors (see Section 3.2 for more details). The output of SCA is an embedding tensor consisting of far-end reference information and near-end microphone information and the input of the encoder. Each convolution block of the encoder is built by a gated convolutional layer [10] which includes a gated linear unit (GLU) activation function and batch-norm layer. The number of kernels are 8, 16, 16 and all convolution kernels have a size of 2×2 with stride of 1×2 and we make the convolution layer causal by padding the incoming signal. At the decoder, each transpose convolution block is composed by a gated transposed convolution layer with GLU activation function followed by a batch norm layer. The number of kernels at the encoder are 16, 8, 2 and all convolution kernels have size 2×2. At the end the decoder, complex projection layer and gate mask layer are applied. The estimated mask is applied to the near-end microphone signal. Each encoder and corresponding decoder layers are connected with a skip connection.

2.3. Streaming cross-attention

Assume that the complex projection outputs from near-end mic recording $d(n)$ and far-end reference input $x(n)$ are $[l_r, l_i]$ and $[f_r, f_i]$, respectively. The subscript r denotes the real part and i the imaginary part. Both $[l_r, l_i]$ and $[f_r, f_i]$ are stored as $\mathbb{R}^{b,t,c,d}$ where b is the batch size, t is the sequence length, c is the 2 channels which are real channel and imaginary channel, and d is the output dimension.

The cross-attention uses a shared multiheads self-attention [9] to explore the long term dependencies for both real and imaginary part. Given a pair input l_i and f_i, Fig 1(b) illustrates the way to get the cross attention output $a_{l,f,i}$. The detailed formulations are as follows:

$$Q = \text{LayerNorm}(f_i), \tag{2}$$
$$K = \text{LayerNorm}(l_i), \tag{3}$$
$$V = \text{LayerNorm}(l_i). \tag{4}$$

Based on the Q, K, V, the multihead attention[9] is applied

$$\text{Attention}(q,k,v) = \text{Softmax}(\frac{\text{Mask}(qk^T)}{\sqrt{d_h}})v, \quad (5)$$

$$head_n = \text{Attention}(QW_n^Q, KW_n^K, VW_n^V), \quad (6)$$

$$\text{MHA}(Q,K,V) = \text{Concate}(head_1, ..., head_h)W^O, \quad (7)$$

where d_h is the dimension for each attention head. Assume the number of heads is N set as 4 in this work, then $d_h = \frac{d}{N}$. $W^O \in \mathbb{R}^{d,d}$ is the projection matrix for the attention output projection. $W_n^Q \in \mathbb{R}^{d,d_h}$, $W_n^K \in \mathbb{R}^{d,d_h}$, $W_n^V \in \mathbb{R}^{d,d_h}$ are the projection matrix for Q, K and V in each attention head, respectively. Similar to [8], the attention mask is used to limit the look-ahead context access for streaming, $\text{Mask}(qk^T)$ masks the product from query and key to be negative infinity, which essentially makes the attention weight to be zero after normalization from softmax.

The cross attention output $a_{l,f,i}$ is result of feeding the multihead attention output through a residual connection, a feed forward network (FFN) and a layer norm operation as follows:

$$a_{l,f,i} = \text{LayerNorm}(\text{FFN}(l_i + \text{MHA}(Q,K,V))). \quad (8)$$

Similarly, we can get $a_{f,l,i}$ the cross attention between f_i and l_i, $a_{f,l,r}$ the cross attention between f_r and l_r, and $a_{l,f,r}$ the cross attention between l_r and f_r. The outputs of the cross attention $[ca_r, ca_i]$ also store in shape of $\mathbb{R}^{b,t,c,d}$ containing two channels from real ca_r and imaginary ca_i. Both ca_r and ca_i are the concatenation of the cross attention output from the real part and imaginary part.

$$ca_r = \text{Concate}(a_{l,f,r}, a_{f,l,r}) \quad (9)$$

$$ca_i = \text{Concate}(a_{l,f,i}, a_{f,l,i}). \quad (10)$$

2.4. Loss function

We train the network with Mean squared error (MSE) loss on time domain and weighed MSE spectral loss on the magnitude spectrum. In the training stage, the enhanced near-end speech signal $\hat{s}(n)$ and target signal $s(n)$ are fed into the MSE loss function. Further, their complex spectrum of these signal processed by STFT and multiplied with a weighting factor are fed into the weighted MSE spectral loss. Formally, the loss function is given by

$$\mathcal{L} = \alpha \sum_n |\hat{s}(n) - s(n)| + \beta \sum_{n,k} w_k |S(\hat{s}(n)) - S(s(n))| \quad (11)$$

where the weighting factors α, β and w_k are heuristically determined to account for distortions in both the low and high frequency regions of the spectrum. The time and frequency indices as n and k for brevity and $S(.)$ is STFT function. In our work, we set $n = 4$, $w_k \in (0.1, 1.0, 1.5, 1.5)$, $\alpha = 1$ and $\beta = \sqrt{512}$.

3. EXPERIMENTATION RESULTS

3.1. Dataset and augmentation

We choose both the synthetic data from AEC-challenge [11] and our private augmented data to train the models. We balance the speakers' genders at both far-end and near-end sides and form total 720 original conversations with each 10s duration. The following typical use cases are considered to augment each conversation.

Reverberation time (RT60) the image method [12] is used to produce both steady and time-variant room impulse responses (RIR) for creating echo paths in typical laptop settings. The RT60 is chosen to have probabilities of 0.6, 0.3, 0.08 and 0.02 over 50 $\sim$ 300 ms, 300 $\sim$ 600 ms, 600 $\sim$ 1 s and 1 $\sim$ 1.5 s. **Delay** between the playback and its received echo is introduced with probabilities of 0.05, 0.6, 0.4, 0.05 over -20 $\sim$ 0 ms, 0 $\sim$ 200 ms, 200 $\sim$ 400 ms and 400 $\sim$ 600 ms. **Signal-to-noise ratio (SNR)** is simulated by using typical noises from DNS-challenge [13] with probabilities of 0.1, 0.1, 0.3 and 0.5 over 0 $\sim$ 10 dB, 10 $\sim$ 20 dB, 20 $\sim$ 30 dB and 30 $\sim$ 40 dB; **Signal-to-echo ratio (SER)** is simulated with probability of 0.1, 0.5, 0.3 and 0.1 over -10$\sim$ 0 dB, 0 $\sim$ 10 dB, 10 $\sim$ 30 dB, 30 $\sim$ 40 dB;**Non-linearity** is simply modelled by either a arc-tangent to imitate gain saturation or a polynomial function as illustrated by [14]; **Time-variant delay/RIR changes** randomly cuts or adds speech and silence segments of 10 to 200 ms to either near-end or far-end signals with a probability of 0 $\sim$ 10%. Time-variant RIR changes are also introduced when generating the echo component. Each augmented conversation further converts to far-end single talk (FEST), near-end single talk (NEST) and double talk (DT) scenarios. The augmentation results in a total of 720K augmented conversations of roughly 2k hours.

3.2. Ablation study

Table 1 illustrates the performance of candidate models using non-streaming (model 1$\sim$4) and streaming (model 5-6) manners over the augmented evaluation data. The non-streaming model would use the whole recording information for AEC, which is suitable for offline AEC task such as video editing. The streaming model would only use the previous frames to remove echo in current frame, for example, in video conference. The input of CRN-2, NCA-CRN-4 and CRN-5 has been aligned by generalised cross-correlation (GCC) [15]. Non-streaming cross-attention alignment (NCA) removes the mask $\text{Mask}(qk^T)$ of SCA to support non-streaming model. Compared with CRN-1, CRN-2 shows that the aligned inputs dramatically improve the AEC performance. With the non-streaming cross-attention mechanism, NCA-CRN-3 achieves significant improvement with the echo return loss enhancement (ERLE) and PESQ [16] increased by 20.7% and 23.2% respectively, as compared to CRN-2. NCA-CRN-4 shows that the performance remains almost the same with additional of-

id	Model	Aligned Input	Attention	RNN	Size	ERLE of FEST	PESQ of NEST	PESQ of DT
1	CRN	No	No	BLSTM	7.8M	24.32	4.29	1.82
2	CRN	Global GCC	No	BLSTM	7.8M	33.27	4.55	2.46
3	NCA-CRN	No	NCA	BLSTM	7.8M	40.17	4.54	3.03
4	NCA-CRN	Global GCC	NCA	BLSTM	7.8M	40.20	4.55	3.02
5	CRN	Streaming GCC	No	LCBLSTM	7.8M	23.68	4.36	2.01
6	SCA-CRN	No	SCA	LCBLSTM	7.8M	32.17	4.50	2.60
7	CRN	Streaming GCC	No	BLSTM	7.8M	27.45	4.42	2.24

Table 1. Performance comparison over candidate models. NCA: Non-streaming cross-attention alignment. SCA: Streaming cross-attention alignment. We measure WB-PESQ both DT and NEST scenarios and ERLE for FEST scenario in the augmented evaluation dataset. The unit for ERLE is db.

	Delay					
Model	50ms	100ms	200ms	300ms	400ms	600ms
CRN-5	22.46	22.78	23.35	21.56	11.52	0.02
NCA-CRN	32.48	32.13	30.76	30.51	34.22	7.34
SCA-CRN	27.49	26.15	25.00	22.72	12.67	2.47

Table 2. Comparison of unaligned modeling performance for FEST scenario vs delays.

	Interspeech 2021		ICASSP 2022
Model	FEST	DT	All
Align-CRUSE	4.46	4.56	N/A
GT-CrossNet	N/A	N/A	4.29
NCA-CRN	4.55	4.69	4.29
SCA-CRN	4.50	4.67	4.27

Table 3. AECMOS comparison against Align-CRUSE and GT-CrossNet baseline approaches.

fline alignment introduced to NCA-CRN-3. Those evidences verify that NCA-CRN-4 is capable of replacing offline alignment as desired. CRN-5 and SCA-CRN are modified versions to support streaming processing without looking ahead. With the same model size, SCA-CRN improves ERLE and PESQ by 36% and 29% respectively. We also test the model with the streaming GCC and BLSTM (model 7). There is 5.8db ERLE decrement if we only place Global GCC by Streaming GCC, becase the delay and path of the ehco is dynamically time-varying. To illustrate that how cross-attention alignment handles data with different delays, we select a subset from the evaluation data where only delay variation and FEST scenario are considered. We choose CRN-5 as baseline to compare with both NCA-CRN and SCA-CRN over ERLE performance with respect to the delays. Table 2 shows that NCA-CRN suppresses the most echo with 5$\sim$20dB ERLE improvement from the baseline over the delay up to 400ms. SCA-CRN suppresses additional 1$\sim$5dB from the baseline within 400ms delay, which covers most of the normal use cases in voice communication. We exclude the non-causal delays in the comparison since SCA-CRN processes only previous data.

3.3. Comparison with the state-of-the-art methods

We use AECMOS, a non-intrusive model-based metric provided by AEC challenge, to compare our approaches with the chosen baselines Align-CRUSE [3] and GT-CrossNet [17]. Align-CRUSE [3] is the SOTA end-to-end AEC network with alignment layer. GT-CrossNet [17] won the runner-up in AEC Challenge 2022 and the detail shown in their paper would help us to compare with our method. The blind test data from AEC challenge ICASSP2022 [11] is used to compare against GT-CrossNet and the counterpart from Interspeech2021 [18] is used to compare against Align-CRUSE. Both data sets are real-world and the realistic distribution of delays has been demonstrated in [3]. Besides FEST and DT scenarios, in "All" scenario, we take the average of all 4 types - FEST echo MOS, NEST other MOS, DT other and echo MOS to indicate the overall performance. Table 3 shows both NCA-CRN and SCA-CRN outperform Align-CRUSE in FEST and DT scenarios and are on par with GT-CrossNet in "All" scenario. From complexity perspective, SCA-CRN is less than half the model size of GT-CrossNet (17.4M) for data of 48k sampling frequency. Due to "NEST other MOS" of Align-CRUSE is not available [3], we can not calculate the "All" scenario and show "N/A" in the table.

4. CONCLUSION

We proposed a novel streaming cross-attention and apply this attention on the convolution recurrent echo cancellation network. With multi-head cross-attention, layer norm, FFN layer and projection operations, the proposed SCA-CRN is able to handle unaligned input signals as well as other challenging echo scenarios, such as time-variant echo path, without additional costly alignment processing. The non-streaming version of SCA-CRN, named as NCA-CRN is proposed for non-streaming echo cancellation task. SCA-CRN and NCA-CRN both achieve encouraging improvement over real public data as compared with baseline approaches, especially in double talk scenarios. In future work, SCA can also be used to address multi-microphone alignment for speech enhancement.

5. REFERENCES

[1] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2021, p. 2472–2476.

[2] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *Interspeech*, 2019, pp. 4255–4259.

[3] Evgenii Indenbom, Nicolae-Cătălin Ristea, Ando Saabas, Tanel Pärnamaa, and Jegor Gužvin, "Deep model with built-in self-attention alignment for acoustic echo cancellation," *arXiv preprint arXiv:2208.11308*, 2022.

[4] Ziteng Wang, Yueyue Na, Zhang Liu, Biao Tian, and Qiang Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the AEC-challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 141–145.

[5] Deepak Kumar Gupta, Vijay Kumar Gupta, and Mahesh Chandra, "Review paper on linear and nonlinear acoustic echo cancellation," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*. Springer, 2015, pp. 465–473.

[6] Hao Zhang, Srivatsan Kandadai, Harsha Rao, Minje Kim, Tarun Pruthi, and Trausti Kristjansson, "Deep adaptive aec: Hybrid of deep learning and adaptive acoustic echo cancellation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 756–760.

[7] J Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[8] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6783–6787.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Conference and Workshop on Neural Information Processing Systems*, 2017.

[10] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[11] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, and Robert Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, 2022.

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

[13] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sergiy Matusevych, Sebastian Braun, Emre Sefik Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner, "ICASSP 2022 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, 2022.

[14] Chenggang Zhang, Jinjiang Liu, and Xueliang Zhang, "LCSM: A lightweight complex spectral mapping framework for stereophonic acoustic echo cancellation," *arXiv preprint arXiv:2208.07277*, 2022.

[15] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 509–519, 2004.

[16] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[17] Haoran Zhao, Nan Li, Runqiang Han, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9112–9116.

[18] Ross Cutler, Ando Saabas, Tanel Pärnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, Robert Aichner, et al., "INTERSPEECH 2021 acoustic echo cancellation challenge.," in *Interspeech*, 2021, pp. 4748–4752.