OFF-THE-SHELF DEEP INTEGRATION FOR RESIDUAL-ECHO SUPPRESSION

Amir Ivry *Israel Cohen* *Baruch Berdugo*

Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering
Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel

ABSTRACT

Residual-echo suppression (RES) systems suppress the echo and preserve the speech from a mixture of the two. In hands-free speech communication, RES may also be addressed as a source separation (SS) or speech enhancement (SE) problem, where the echo can be manipulated as an interfering speech signal. In this study, we fine-tune three pre-trained deep learning-based systems originally designed for RES, SS, and SE, and show that the best performing system for the task of RES varies with respect to the acoustic conditions. Then, we propose a real-time data-driven integration of these systems, where a neural network continuously tracks the system that achieves the best performance during both single-talk and double-talk periods. Experiments with 100 h of real and synthetic data show that the integrated system outperforms each individual system in terms of echo suppression and speech distortion in various acoustic environments.

***Index Terms*—** Acoustic echo cancellation, residual-echo suppression, speech separation, speech enhancement, deep learning.

1. INTRODUCTION

Hands-free speech communication often involves a conversation between two speakers located at near-end and far-end points. The near-end microphone captures the desired-speech signal, echo produced by a loudspeaker playing the far-end signal, and background noise. The acoustic coupling between the loudspeaker and the microphone may lead to degraded speech intelligibility in the far-end due to echo presence [1]. Numerous acoustic echo cancellation (AEC) systems were proposed to reduce the echo of the far-end speaker's speech and preserve the near-end speaker [2]. However, echos are often not eliminated by AEC systems and must be further reduced using residual-echo suppression (RES) systems.

RES can also be addressed as a speech separation (SS) [3] or speech enhancement (SE) [4] problem, where the echo is considered an interfering speech signal. In this study, we first fine-tune three off-the-shelf deep-learning-based systems: Our recently introduced RES system [5], a convolutional time-domain audio separation network (Conv-TasNet) [6], and a denoiser develop by Facebook™ for SE [7]. We show that the best-performing system of the three varies depending on the speech, echo, and noise levels. Second, we propose a real-time data-driven integration of these systems using a deep neural network (NN) that continuously tracks the best system based on single-talk and double-talk performance measures. Experiments with 100 h of real and synthetic data show that the integrated system achieves better performance than each system in terms of echo cancellation and speech distortion across various acoustic setups in both single-talk and double-talk.

This research was supported by the Pazy Research Foundation.

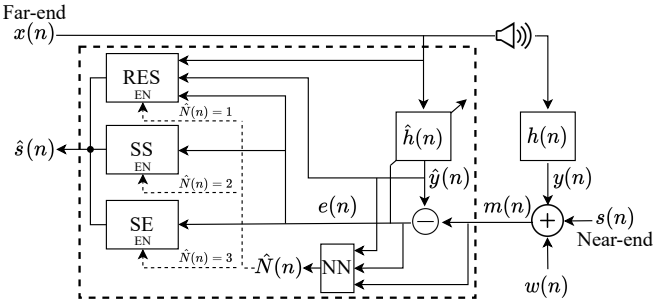

Fig. 1: AEC scenario and proposed system integration.

2. PROBLEM FORMULATION

Figure 1 depicts the AEC scenario and proposed system. Let $s\,(n)$ be the near-end speech signal and let $x\,(n)$ be the far-end speech signal, where n is the time index. The microphone signal $m\,(n)$ is given by $m\,(n) = s\,(n) + y\,(n) + w\,(n)$, where $w\,(n)$ represents additive environmental and system noises and $y\,(n)$ is a nonlinear and reverberant echo that is generated from $x\,(n)$. First, an AEC system receives $m\,(n)$ as input and $x\,(n)$ as reference, and generates two signals: the echo estimate $\hat{y}\,(n)$, and the near-end signal estimate $e\,(n)$ given by:

$$\begin{aligned} e\,(n) &= m\,(n) - \hat{y}\,(n) \\ &= s\,(n) + (y\,(n) - \hat{y}\,(n)) + w\,(n)\,. \end{aligned} \tag{1}$$

A succeeding system aims to cancel the residual echo by eliminating $y\,(n) - \hat{y}\,(n)$, without distorting the speech $s\,(n)$. The NN continuously selects and enables the best out of the RES, SS, and SE systems that interchangeably perform this task.

3. PROPOSED DEEP INTEGRATION SYSTEM

We consider three systems that were originally constructed and pre-trained for RES, SS, and SE. For RES, we employ an extension of the system in [5]. It comprises a U-net [8] NN that is fed with the short-time Fourier transform (STFT) [9] amplitude of $e(n)$, $\hat{y}(n)$, and $x(n)$, and aims to recover the STFT amplitude of $s(n)$. The objective function that is minimized during training is the mean squared error between the NN prediction and $s(n)$. The employed SS system is the waveform-based Conv-TasNet [6]. It comprises an encoder that maps the error mixture $e(n)$ to a high-dimensional representation, and a separation module that calculates a mask for each speech source in the mixture, i.e., the near-end speech and echo. Then, a decoder reconstructs the desired source from the masked features. A 1-D convolutional autoencoder [10] models the waveforms, and a temporal convolutional network separation module [11] estimates the masks based on the encoder output. The scale-invariant source-to-noise ratio [12] is maximized during optimization, which is a modified version of the standard signal-to-distortion ratio [13]. The SE system that is applied is the waveform-based NN in [7] that receives $e(n)$ and aims to cancel the residual echo and noise from it. The proposed model is based on an encoder-decoder architecture with skip-connections [14]. It is optimized on both time and frequency domains using multiple loss functions. Namely, the ℓ_1 loss over the waveform together with a multi-resolution STFT loss over the spectrogram magnitudes are jointly minimized.

The proposed integrated system includes a deep NN that receives the waveform representations of $e(n)$, $\hat{y}(n)$, and $m(n)$, and finds the best out of the RES, SS, and SE systems. The training stage of the NN is done as follows. First, all three pre-trained systems are fine-tuned separately and independently with an identical training database. Then, a validation set is propagated via each fine-tuned system, and two performance measures are extracted from each system. During single-talk periods, the echo return loss enhancement (ERLE) [15] is used. It measures echo reduction between the degraded and enhanced signals when only a far-end signal and noise are present and is given by $10\log_{10}\left[\|e(n)\|_2^2/\|\hat{s}(n)\|_2^2\right]$. During double-talk, the objective deep noise suppression mean opinion score (DNSMOS) metric is used [16], which estimates objective human ratings. In [17], the DNSMOS has shown a strong correlation with echo suppression and speech preservation measures for the task of RES during double-talk. These measures are used to form a second training set as follows. Every time frame in the validation set is attached to a new categorical label, $N(n)$, from the set $\{1, 2, 3\}$, corresponding to the RES, SS, and SE systems. $N(n)$ is assigned to the index of the system with the highest ERLE during single-talk or highest DNSMOS during double-talk. This new dataset is used for training the NN. The NN architecture is waveform-based and follows the one in [18]. Still, its input layer is extended to three channels instead of two, and its final layer is concatenated to an additional softmax layer with three output neurons. In real-time, unseen data are propagated via the NN that yields the index estimate of the best system, denoted by $\hat{N}(n)$, and the respective fine-tuned system is enabled to execute RES.

The proposed NN contains 19 thousand parameters that consume 520 Mflops and 42 KB of memory. Thus, its integration on hands-free devices is enabled, e.g., using the NDP120 neural processor by Syntiant™ [19]. Timing constraints of hands-free communication on that processor are also met [20]. The preceding AEC reduces linear echo with a standard normalized least mean squares (NLMS) [21] adaptive filter with a filter length of 150 ms and step size of 3×10^{-5}.

4. EXPERIMENTAL SETUP

The AEC challenge database [22] is employed in this study. This corpus is sampled at 16 kHz and includes single-talk and double-talk periods, both with and without echo-path change. No echo-path change means no movement in the room during the recording, and echo-path change means either the near-end speaker or the device is moving during the recording. The corpus includes 25 h of synthetic data and 75 h of real clean and noisy data. The speech-to-echo ratio (SER), speech-to-noise ratio (SNR), and echo-to-noise ratio (ENR) levels were distributed on $[-20, 10]$ dB, $[0, 40]$ dB, and $[0, 40]$ dB, respectively. These ratios are defined as:

$$\text{SER} = 10\log_{10}\left[\|s(n)\|_2^2/\|y(n)\|_2^2\right], \quad (2)$$

$$\text{SNR} = 10\log_{10}\left[\|s(n)\|_2^2/\|w(n)\|_2^2\right], \quad (3)$$

$$\text{ENR} = 10\log_{10}\left[\|y(n)\|_2^2/\|w(n)\|_2^2\right], \quad (4)$$

and calculated with 50% overlapping time frames of 20 ms.

The 100 h of data is randomly split to create 80 h of training, 10 h of validation, and 10 h of test sets. Each set is divided into 10 s segments that contain recordings in different setups. This leads to frequent re-convergence during transitions between segments, both without and with echo-path change. All sets are balanced to prevent a bias in the results, as described in [5]. During fine-tuning, each system maintains its original training configurations, but with an initial learning rate of 10 times smaller. For the NN, the data pre-processing follows [18], and the NN is trained with back-propagation through time with a learning rate of 5×10^{-4}, mini-batch of 40 ms, and 20 epochs, using Adam optimizer [23]. The minimized objective function is the categorical cross-entropy [24] between the prediction and one-hot-vector encoding [25] of the optimal-system index $N(n)$. Training duration was typically 15 minutes per 10 h of data, and inference time was 12 ms per batch on an Intel Core i7-8700K CPU @ 3.7 GHz with two GPUs of type Nvidia GeForce RTX 2080 Ti.

Table 1: Performance with no echo-path change.

	INT	RES	SS	SE
DNSMOS	**3.1±0.8**	2.4±0.5	2.5±0.8	2.6±1.0
DSML	**9.6±1.0**	8.7±0.8	9.1±1.1	9.2±1.2
RESL	**29.6±4.5**	27.5±3.5	28.3±4.3	28.5±4.6
ERLE	**33.1±1.6**	32.5±2.0	32.2±1.7	32.1±1.4

Table 2: Performance with echo-path change.

	INT	RES	SS	SE
DNSMOS	**2.4±0.5**	1.8±0.3	2.0±0.6	2.1±0.6
DSML	**9.2±0.7**	8.4±0.6	8.8±0.7	8.9±0.8
RESL	**27.0±3.0**	25.3±2.5	25.7±3.0	25.8±3.2
ERLE	**30.5±2.2**	28.5±2.8	28.2±2.3	28.3±1.8

Table 3: Performance before linear AEC convergence.

	INT	RES	SS	SE
DNSMOS	**2.1±0.2**	1.7±0.2	1.8±0.3	1.9±0.3
DSML	**7.6±1.1**	7.1±0.7	7.2±0.9	7.4±1.1
RESL	**24.2±4.4**	22.5±3.0	22.8±3.5	23.0±4.4
ERLE	**27.7±2.2**	26.0±3.2	25.8±2.8	25.4±2.0

To separately measure echo suppression and speech distortion in double-talk, we respectively employ the recently introduced residual-echo suppression level (RESL) and desired-speech maintained level (DSML) [17]:

$$\text{RESL} = 10\log_{10}\left[\|r\left(n\right)\|_2^2/\|g\left(n\right)r\left(n\right)\|_2^2\right], \quad (5)$$

$$\text{DSML} = 10\log_{10}\left[\|\tilde{s}\left(n\right)\|_2^2/\|\tilde{s}\left(n\right)-g\left(n\right)s\left(n\right)\|_2^2\right], \quad (6)$$

where $g\left(n\right) = \widehat{s}\left(n\right)/e\left(n\right)$ is the time varying gain of the NN, $r\left(n\right) = e\left(n\right) - s\left(n\right)$ is the aligned noisy echo estimate, and $\tilde{s}\left(n\right) = \widehat{g}\left(n\right)s\left(n\right)$, where

$$\widehat{g}\left(n\right) = \left\langle g\left(n\right)s\left(n\right), s\left(n\right)\right\rangle/\|s\left(n\right)\|_2^2. \quad (7)$$

The DNSMOS [16] is used again during double-talk to assess speech quality for human perception. During single-talk, the echo suppression level is quantified using the ERLE [15]. The DSML, RESL, and ERLE are calculated with 50% overlapping time frames of 20 ms, and the DNSMOS is applied with the API provided by Microsoft™.

5. EXPERIMENTAL RESULTS

The integrated system, denoted as "INT", is compared against the particular RES, SS, and SE systems. In Tables 1–3, performance measures are given with their mean and standard deviation (std) values in the format mean±std. In Figures 2–4, only the average values of the performance measures are shown. For all the measures, higher mean and lower std indicate better performance. Convergence of the linear AEC is assumed if the normalized misalignment was lower than -10 dB for a given echo path [21]. The results are derived with respect to the entire test set.

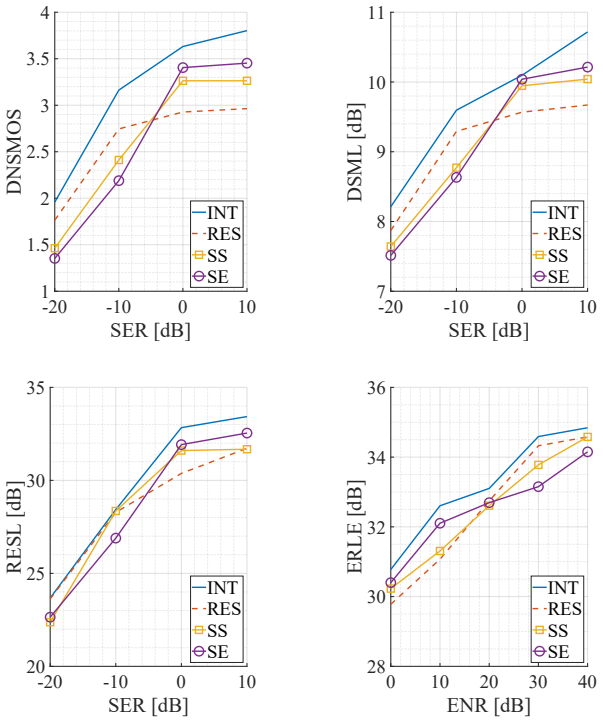

Fig. 2: Comparison of the integrated system and individual systems for segments with no echo-path change.

Results for no echo-path change are given in Table 1, and for echo-path change are shown in Table 2, both after convergence. In Table 3, results for no echo-path change before convergence are reported. Comparing the RES, SS, and SE systems, we may conclude that the SE system obtains better average performance during double-talk periods in terms of echo cancellation as shown by the RESL, desired-speech distortion as shown by the DSML, and speech quality as demonstrated by the DNSMOS. However, the SE system also obtains the highest std values in double-talk, making it less stable than competition. The RES system is favorable in single-talk echo cancellation with a higher average ERLE, but is also the least stable with the highest std value. These observations also hold for echo-path change and pre-convergence scenarios, but with an expected degradation in the values of all performance measures. Thus, neither the RES, SS, nor SE system is optimal across all measures and acoustic scenarios in terms of higher average performance and in terms of lower std values. The proposed integrated system outperforms each individual system on average across all measures and scenarios during both single-talk and double-talk periods.

Average performance is also analyzed for various levels of SER during double-talk and multiple levels of ENR during

single-talk. Results for segments without and with echo-path change are given in Figures 2 and 3, respectively, both after convergence. Results for segments with no echo-path change before convergence are shown in Figure 4. During double-talk, the SE system outperforms the RES and SS systems when SER levels are high, and the RES system is preferable when SER levels are low. During single-talk, the RES system obtains higher performance when ENR levels are high, and the SE system is preferable for low levels of ENR. These observations remain across all measures and also for echo-path change and pre-convergence scenarios, but again with an expected overall decrease on the average performance. These results reaffirm that the best performing system varies with speech, echo, and noise levels, and supports previous claims that no individual system can be considered best under all acoustic conditions.

A possible explanation for the behavior of the three systems with respect to acoustic conditions is suggested. The SE system is better suited to handle high SER levels since the noisy echo is significantly attenuated with respect to speech and appears as a noisy interference. Similarly, as ENR decreases, the SE system is successful since the echo is mainly screened by noise. The RES system is preferable when the SER level is low, since it was trained to detect residual-echo signatures that are mixed with speech. Likewise, when ENR levels are high, the residual echo dominates the signal and can be successfully recognized and suppressed by the RES system. The proposed integrated system outperforms the RES, SS, and SE systems across all speech, echo, and noise levels, in both no echo-path change, echo-path change, and pre-convergence scenarios. Based on the presented results, it can be concluded that the proposed NN can estimate which system is best in real-time for various acoustic conditions during both single-talk and double-talk periods, and that the integrated system is better on average than each of its three components.

6. CONCLUSIONS

We have introduced a real-time data-driven system integration framework and applied it to the task of RES. This integration comprises three deep learning-based systems originally constructed and pre-trained for RES, SS, and SE. After fine-tuning all three systems and showing that none of these systems can be considered best for RES, we developed a deep NN that continuously selects the best of the RES, SS, and SE systems and enables it to perform RES. Using 100 h of real and synthetic recordings, we showed that the NN can estimate the best system in real time and that the proposed integrated system outperforms, on average, each of the three individual systems in terms of echo cancellation and speech distortion during both single-talk and double-talk periods.

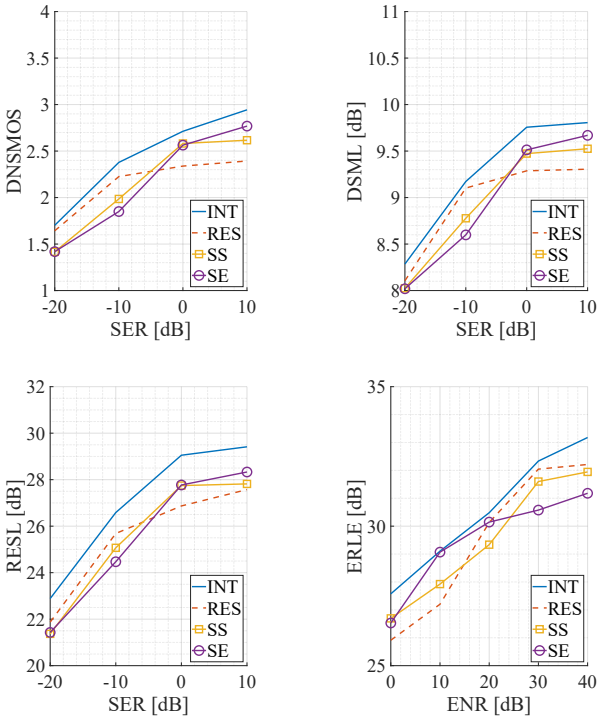

Fig. 3: Comparison of the integrated system and individual systems for segments with echo-path change.

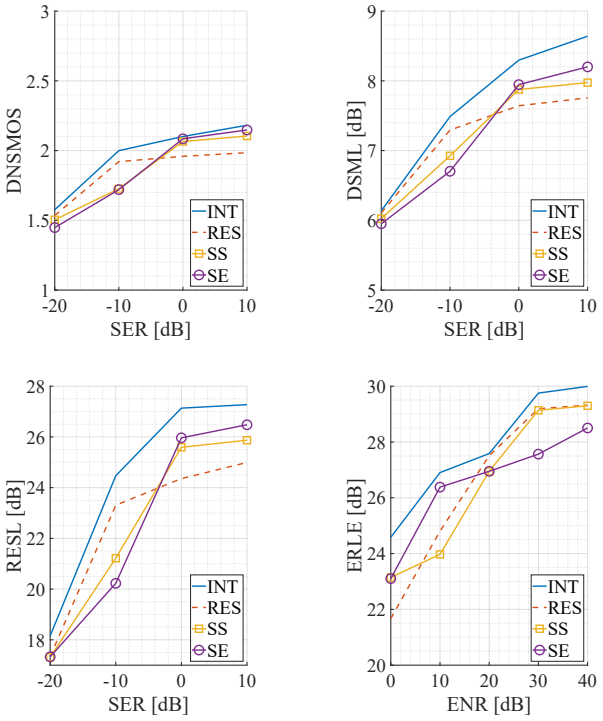

Fig. 4: Comparison of the integrated system and individual systems for segments before the linear AEC convergence.

7. REFERENCES

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, 1995.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, et al., *Advances in Network and Acoustic Echo Cancellation*, New York: Springer, 2001.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2014.

[5] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *Proc. ICASSP*. IEEE, 2021, pp. 126–130.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *preprint arXiv:2006.12847*, 2020.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. on Med. Image Comput. and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[10] K. G. Ghasedi, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proc. ICCV*, 2017, pp. 5736–5745.

[11] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. ECCV*. Springer, 2016, pp. 47–54.

[12] M. Chao, L. Dongmei, and J. Xupeng, "Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment," in *Proc. APSIPA ASC*. IEEE, 2020, pp. 711–715.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] D. Michal, V. Eugene, C. Gabriel, K. Samuel, and P. Chris, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187. Springer, 2016.

[15] "ITU-T Rec. G.168: Digital network echo cancellers," Feb. 2012.

[16] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*. IEEE, 2021, pp. 6493–6497.

[17] A. Ivry, I. Cohen, and B. Berdugo, "Objective metrics to evaluate residual-echo suppression during double-talk," in *Proc. WASPAA*, 2021.

[18] A. Ivry, I. Cohen, and B. Berdugo, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. Interspeech*, 2021, pp. 4773–4777.

[19] "NDP120 Syntiant™ Neural Processor," `https://www.syntiant.com/ndp120`, 2021.

[20] "ETSI ES 202 740: Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user," 2016.

[21] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *EURASIP J. on Adv. in Signal Process.*, vol. 97, no. 1, pp. 1–19, 2015.

[22] R. Cutler, A. Saabas, T. Parnamaa, M. Loida, S. Sootla, et al., "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*, 2021, pp. 4748–4752.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[24] Z. Zhilu and R. S. Mert, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. NeurIPS*, 2018, p. 8792–8802.

[25] P. Kedar, S. P. Taher, and D. P. Chinmay, "A comparative study of categorical variable encoding techniques for neural network classifiers," *Int. J. of Comput. Appl.*, vol. 175, no. 4, pp. 7–9, 2017.