OBJECTIVE METRICS TO EVALUATE RESIDUAL-ECHO SUPPRESSION DURING DOUBLE-TALK

Amir Ivry *Israel Cohen* *Baruch Berdugo*

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

2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 978-1-6654-4870-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/WASPAA52581.2021.9632787

ABSTRACT

Human subjective evaluation is optimal to assess speech quality for human perception. The recently introduced deep noise suppression mean opinion score (DNSMOS) metric was shown to estimate human ratings with great accuracy. The signal-to-distortion ratio (SDR) metric is widely used to evaluate residual-echo suppression (RES) systems by estimating speech quality during double-talk. However, since the SDR is affected by both speech distortion and residual-echo presence, it does not correlate well with human ratings according to the DNSMOS. To address that, we introduce two objective metrics to separately quantify the desired-speech maintained level (DSML) and residual-echo suppression level (RESL) during double-talk. These metrics are evaluated using a deep learning-based RES-system with a tunable design parameter. Using 280 hours of real and simulated recordings, we show that the DSML and RESL correlate well with the DNSMOS with high generalization to various setups. Also, we empirically investigate the relation between tuning the RES-system design parameter and the DSML-RESL tradeoff it creates and offer a practical design scheme for dynamic system requirements.

Index Terms— Residual-echo suppression, echo cancellation, objective metrics, perceptual speech quality, deep learning.

1. INTRODUCTION

Hands-free communication often involves a conversation between two speakers located at near-end and far-end points. The near-end microphone can capture the desired-speech signal and two interfering signals: nonlinear echo produced by a loudspeaker playing the far-end signal, and background noises [1, 2]. The acoustic coupling between the loudspeaker output and the microphone may lead to degraded speech intelligibility in the far-end due to echo presence [3]. The most challenging scenarios are double-talk periods, when the desired speech and echo are captured by the microphone at the same time. To combat that, numerous nonlinear acoustic echo cancellation (NLAEC) systems were proposed to remove the nonlinear echo and to preserve the near-end speech [4–8]. However, often there is still a mismatch between true and estimated echo paths, especially during the NLAEC convergence and re-convergence [9, 10]. As a result, the echo is not eliminated and the NLAEC should be followed by a residual-echo suppression (RES) system.

Human perception of speech quality is optimally evaluated using human subjective evaluation [11]. Lately, the objective deep noise suppression mean opinion score (DNSMOS) metric has been proposed to estimate human ratings and has shown great accuracy [12]. Regarding the task of RES, speech quality during double-talk is traditionally evaluated using the objective signal-to-distortion ratio (SDR) metric [13], e.g., in [14–19]. Unfortunately, the SDR is affected by both desired-speech distortion and residual-echo presence, which renders it unreliable in predicting the DNSMOS and unreliable in predicting human perception of speech quality [12].

This paper introduces two objective metrics that separately evaluate the desired-speech maintained level (DSML) and the residual-echo suppression level (RESL) during double-talk. Considering the RES system as a time-varying gain, the DSML is obtained by applying that gain to the desired speech and substituting the outcome in the definition of the SDR. The RESL is obtained by subtracting the desired speech from the double-talk segment and calculating the ratio of the noisy residual-echo before and after the gain is applied to it. To evaluate these metrics, we employ a deep learning-based RES system that also embeds a design parameter [20]. Experiments are done with 280 h of real and simulated recordings in various scenarios and in high and low levels of echo and noise. Results show that the DSML and RESL have high correlation with human perception according to the DNSMOS, and high generalization to various setups, which renders them more suitable for speech quality evaluation than the SDR. We further investigate the empirical relation between tuning the design parameter and the DSML-RESL tradeoff it creates. Based on this relation, we offer a practical scheme for tuning the design parameter during training to optimally cope with dynamic system requirements.

The remainder of this paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we introduce the DSML and RESL metrics. Section 4 covers the employed RES system and its tunable design parameter. Section 5 describes the database and additional performance metrics, and experimental results are presented in Section 6. Section 7 concludes this study.

2. PROBLEM FORMULATION

Figure 1 depicts the RES scenario. Let $s(n)$ be the desired near-end speech signal and let $x(n)$ be the far-end speech signal. The near-end microphone signal $m(n)$ is given by

$$m(n) = s(n) + y(n) + w(n), \tag{1}$$

where $w(n)$ represents additive environmental and system noises and $y(n)$ is a reverberant echo that is nonlinearly generated from $x(n)$. Before applying RES, the NLAEC system introduced in [8] is applied to reduce nonlinear echo. The NLAEC receives $m(n)$ as input and $x(n)$ as reference, and generates two signals: the echo estimate $\widehat{y}(n)$, and the desired-speech estimate $e(n)$, given by

$$e(n) = m(n) - \widehat{y}(n) = s(n) + [y(n) - \widehat{y}(n)] + w(n). \tag{2}$$

This research was supported by the Pazy Research Foundation and the ISF-NSFC joint research program (grant No. 2514/17). The authors thank Stem Audio for providing equipment and technical guidance.

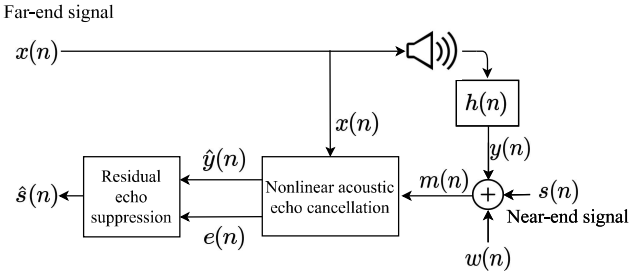


Figure 1: Residual-echo suppression scenario.

The goal of the RES system is to suppress the residual echo $y(n) - \widehat{y}(n)$ without distorting the desired-speech signal $s(n)$.

3. DSML AND RESL

To derive the DSML and RESL, a deep learning-based RES system is considered as a time-varying gain. During double-talk, $e(n) \neq 0$ and the gain is given by

$$g(n) = \left.\frac{\widehat{s}(n)}{e(n)}\right|_{\text{Double-talk}}. \tag{3}$$

Before introducing the DSML and RESL metrics, the SDR and its drawbacks are examined. According to [13], the SDR is defined as

$$\begin{aligned}\text{SDR} &= 10\log_{10}\left.\frac{\|s(n)\|_2^2}{\|s(n)-\widehat{s}(n)\|_2^2}\right|_{\text{Double-talk}} \\ &= 10\log_{10}\left.\frac{\|s(n)\|_2^2}{\|s(n)-g(n)\,e(n)\|_2^2}\right|_{\text{Double-talk}}.\end{aligned} \tag{4}$$

The SDR is affected by both the desired-speech distortion and residual-echo presence, and makes no distinction between cases in which $g(n)\,e(n)$ comprises distortion-free speech and echo, or distorted speech without echo. Thus, the SDR does not correlate well with human ratings [12], since these scenarios clearly exhibit different human perception ratings and different DNSMOS values. A distinction between desired-speech distortion and residual-echo suppression is extremely valuable for evaluating RES during double-talk. Hence, we propose two objective metrics by applying $g(n)$ separately to the desired speech and noisy residual-echo estimate.

Formally, the DSML is calculated similarly to the SDR, but $g(n)$ is applied only to the desired speech $s(n)$:

$$\text{DSML} = 10\log_{10}\left.\frac{\|\tilde{s}(n)\|_2^2}{\|\tilde{s}(n)-g(n)\,s(n)\|_2^2}\right|_{\text{Double-talk}}. \tag{5}$$

The RESL is derived by estimating the noisy residual-echo as $r(n) = e(n) - s(n)$, and evaluating the following ratio:

$$\text{RESL} = 10\log_{10}\left.\frac{\|r(n)\|_2^2}{\|g(n)\,r(n)\|_2^2}\right|_{\text{Double-talk}}. \tag{6}$$

Note that the RES system may introduce a constant attenuation that leads to an artificial desired-speech distortion in the DSML. To ensure the DSML is invariant to that attenuation, it is compensated as in [14]. Explicitly, $\tilde{s}(n) = \widehat{g}(n)\,s(n)$, where:

$$\widehat{g}(n) = \frac{\langle g(n)\,s(n), s(n)\rangle}{\|s(n)\|_2^2}. \tag{7}$$

4. RES SYSTEM WITH A DESIGN PARAMETER

To evaluate the performances of the DSML and RESL, we employ a deep learning-based RES system that embeds a tunable design parameter [20]. This system comprises a UNet neural network [21] with two input channels and one output channel. The network is fed with the short-time Fourier transform (STFT) [22] amplitude of the NLAEC outputs and aims to recover the STFT amplitude of the desired speech. The design parameter $\alpha \geq 0$ is embedded in a custom loss function $J(\alpha)$ that is minimized during training:

$$J(\alpha) = \|\widehat{S}(f) - S(f)\|_2^2 + \alpha\|\widehat{S}(f)\|_2^2 + 0.1\,\sigma^2_{\widehat{S}(f)}\mathbb{I}_{\alpha>0}\,, \tag{8}$$

where $\widehat{S}(f)$ and $S(f)$, respectively, represent the desired-speech prediction and ground truth spectra amplitudes, $\sigma^2_{\widehat{S}(f)}$ denotes the variance of $\widehat{S}(f)$, and $\mathbb{I}_{\alpha>0}$ equals 1 when $\alpha > 0$ and 0 otherwise. During the training stage, $J(\alpha)$ is minimized while α penalizes $\|\widehat{S}(f)\|_2^2$, which allows a dynamic tradeoff between the desired-speech distortion and residual-echo suppression of the system, namely between the DSML and RESL. When $\alpha = 0$, the error between the desired-speech prediction and ground truth is minimized. However, when $\alpha > 0$, smaller prediction values are generated. This reduces the level of residual echo but compromises the level of desired-speech distortion. $\sigma^2_{\widehat{S}(f)}$ mitigates sub-band nullification that may occur when $\alpha > 0$. Note that α and the DSML-RESL tradeoff it creates can be tuned during the training process.

5. EXPERIMENTAL SETUP

5.1. Database

Two data corpora were employed in this study; the AEC-challenge database [23], and a database recorded in our lab, both sampled at 16 kHz. These corpora consider single-talk and double-talk periods both without and with echo-path change. In the former there is no movement during the recording, and in the latter either the near-end speaker or device are moving during the recording. In [23], two open sources of synthetic and real recordings are introduced. The synthetic data includes 100 h, and the real data contains 140 h of audio clips, generated from $5,000$ hands-free devices that are used in various acoustic environments. In both real and synthetic cases, signal-to-echo ratio (SER) and signal-to-noise ratio (SNR) levels were distributed on $[-10, 10]$ dB and $[0, 40]$ dB, respectively. Additional real recordings were conducted in our lab to test the generalization of the DSML and RESL to unseen setups and their robustness to extremely low levels of SERs. This database is fully described in [20]. For completion, it contains 40 h of recordings from the TIMIT [24] and LibriSpeech [25] corpora with SNR levels of 32 ± 5 dB and SER levels distributed on $[-20, -10]$ dB.

The SER is defined as $\text{SER}=10\log_{10}\left[\|s(n)\|_2^2/\|y(n)\|_2^2\right]$ and the SNR is defined as $\text{SNR}=10\log_{10}\left[\|s(n)\|_2^2/\|w(n)\|_2^2\right]$ in dB, each is calculated with 50% overlapping time frames of 20 ms.

5.2. Data Processing, Training, and Testing

The real and synthetic data from [23] was randomly split to create 185 h of training set and 45 h of validation set. The test set contains only real data that includes the remaining 10 h from [23] and all 40 h from [20]. Each set was divided into 10 s segments that contain recordings in different setups. This leads to frequent re-convergence during transitions between segments, both with and

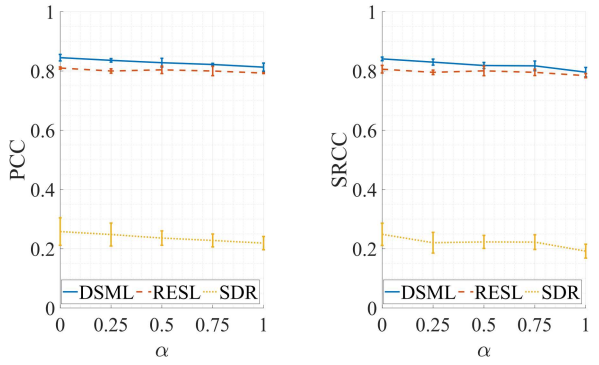


Figure 2: Correlation of DNSMOS with the proposed DSML and RESL metrics, and the widely-used SDR.

without echo-path change. These sets are balanced to prevent bias in the results, as detailed in [20]. The NLAEC system, which is also deep learning-based [8], and the succeeding RES system [20], were trained separately. During testing, in accordance with Section 3, the artificial gain that may be introduced by the RES system is compensated as in [13, 14] before deriving the DSML and RESL.

5.3. Additional Performance Metrics

We employ additional metrics to evaluate RES. The echo return loss enhancement (ERLE) [26] measures echo reduction between the degraded and enhanced signals when only echo and noise are present:

$$\text{ERLE} = 10\log_{10}\left.\frac{\|e(n)\|_2^2}{\|\widehat{s}(n)\|_2^2}\right|_{\text{Far-end single-talk}}. \quad (9)$$

The signal-to-artifacts ratio (SAR) [13] measures the desired-speech distortion during near-end single-talk periods:

$$\text{SAR} = 10\log_{10}\left.\frac{\|s(n)\|_2^2}{\|s(n)-\widehat{s}(n)\|_2^2}\right|_{\text{Near-end single-talk}}. \quad (10)$$

The perceptual evaluation of speech quality (PESQ) [27] metric, which correlates well with the DNSMOS [12], is used in double-talk. The SAR and SDR are compensated as the DSML in eq. (5).

6. EXPERIMENTAL RESULTS

The performance metrics are evaluated using the RES system and are calculated with 50% overlapping frames of 20 ms. The metrics are reported by their mean and standard deviation (std) values in Table 1, and by their mean only in Figures 4 and 5, with respect to the test set specified in each experiment. For all metrics, higher mean and lower std indicate a better performance. In our study, the convergence of the NLAEC follows the definitions in [8,28], and the DNSMOS is calculated using the API provided by Microsoft [12].

First, we explore the correlation of the DSML and RESL with the DNSMOS using Pearson correlation coefficient (PCC) [29] and Spearman's rank correlation coefficient (SRCC) [30], as done in [12, 31]. This experiment includes segments without echo-path change after NLAEC convergence for $\alpha = [0, 0.25, 0.5, 0.75, 1]$, and the results are shown in Figure 2. The conclusion drawn in [12]

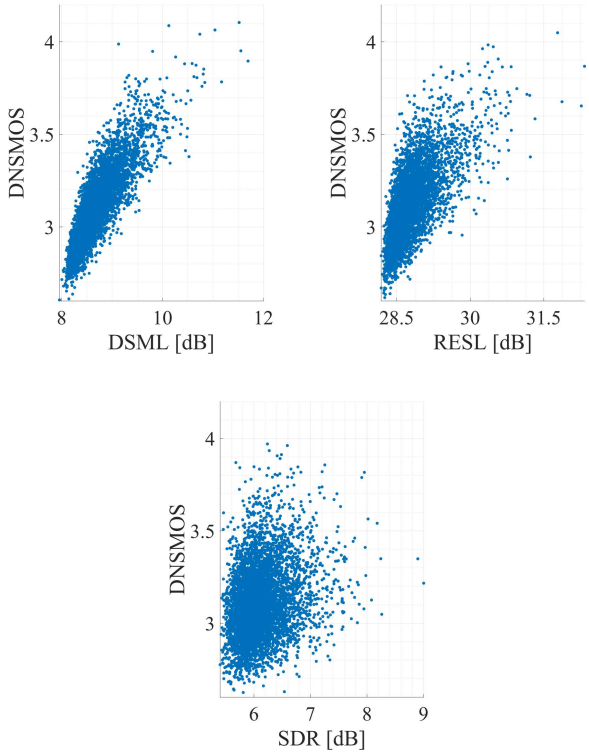


Figure 3: Scatter plots of DNSMOS versus the proposed DSML and RESL metrics, and the widely-used SDR.

is reaffirmed in this study, i.e., the SDR does not correlate well with the DNSMOS, as the PCC and SRCC mean values are below 0.26 for all α. On the contrary, the DSML and RESL are highly correlated with the DNSMOS, with mean correlation scores between 0.78 and 0.85 for all α. Also, compared to the SDR, the DSML and RESL correlations are relatively more consistent across α values, as inferred from their lower std values. To visualize these correlations, Figure 3 depicts scatter plots of the DNSMOS versus the DSML, RESL, and SDR metrics for random sample values with $\alpha = 0$. These plots validate the poor correlation between the DNSMOS and SDR, and the high correlation between the DNSMOS and the DSML and RESL. Conclusively, the DSML and RESL are better correlated with human perception and speech quality evaluation.

All performance metrics are evaluated in Table 1 with $\alpha = 0$. Separate results are shown for segments without and with echo-path change after NLAEC convergence, and for segments before convergence. The DSML and RESL are consistent with all other metrics, which degrade when shifting from no echo-path change to echo-path change scenarios, and further degrade when considering segments before convergence. This also implies high generalization of the DSML and RESL to various setups. The DSML is consistently higher than the SDR, as expected, since the definition in eq. (4) also considers echo and noise in the denominator. Also, the DSML is lower than the SAR, which is applicable to single-talk segments where speech is less distorted by the RES system. The RESL is always lower than the ERLE, which is relevant to segments without desired speech where echo is more suppressed. These observations

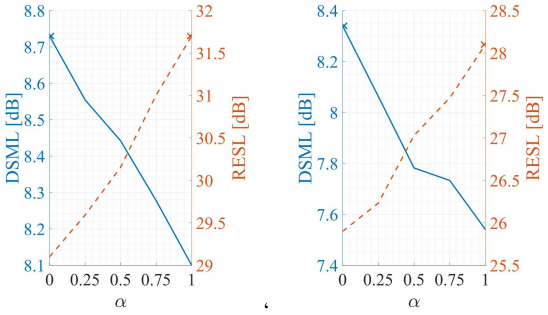
Table 1: Performance metrics in various scenarios with $\alpha = 0$.

	No echo-path change	Echo-path change	Before convergence
DNSMOS	3.12 ± 0.2	2.91 ± 0.3	2.56 ± 0.6
DSML	8.73 ± 0.4	8.34 ± 0.5	6.97 ± 0.7
RESL	29.1 ± 3.7	25.9 ± 4.4	22.1 ± 5.6
SDR	6.13 ± 0.4	5.94 ± 0.6	5.57 ± 0.8
PESQ	3.58 ± 0.2	3.35 ± 0.5	3.18 ± 0.6
SAR	9.88 ± 0.4	9.69 ± 0.5	9.51 ± 0.6
ERLE	33.2 ± 3.1	29.1 ± 4.2	26.4 ± 5.1

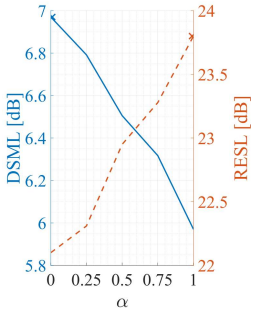
highlight the reliability of the DSML and RESL metrics.

Next, the relation between tuning α and the DSML-RESL tradeoff it creates is investigated. Figure 4 considers segments without and with echo-path change after convergence, and segments before convergence, for $\alpha = [0, 0.25, 0.5, 0.75, 1]$. As α increases, speech is more distorted and the DSML decreases, while residual echo is more suppressed and the RESL increases. This tradeoff occurs across all scenarios and is empirically consistent for all α values. This tradeoff is also analyzed in various SER and SNR levels that occur in real-life setups. In this experiment, segments without echo-path change are considered and results are given in Figure 5. It can be observed that both the DSML and RESL are impaired when acoustic conditions deteriorate, as expected. Also, the relation between α and the metrics is retained, i.e., for all levels of echo and noise, increasing α degrades the DSML and enhances the RESL.

Finally, we offer a practical design scheme for possible dynamic user requirements. Assume an environment without echo-path change after convergence, which can be inferred by the user using the definitions in [8, 28]. At first, the user requires an average RESL higher than 30 dB and DSML higher than 8.4 dB. According to Figure 4(a), $\alpha = 0.5$ is selected. Next, the user evaluates that SER $= 0$ dB and SNR $= 20$ dB, e.g., by respectively analyzing double-talk and near-end single-talk periods, and accordingly decides to suppress the maximal amount of echo that maintains DSML no lower than 8.3 dB. Then, according to Figure 5, the user shifts $\alpha = 0.5$ to $\alpha = 0.75$ during training, which decreases the average DSML to 8.3 dB and increases the average RESL to above 31 dB.



(a) No echo-path change (b) Echo-path change



(c) Before convergence

Figure 4: DSML-RESL tradeoff for various values of α.

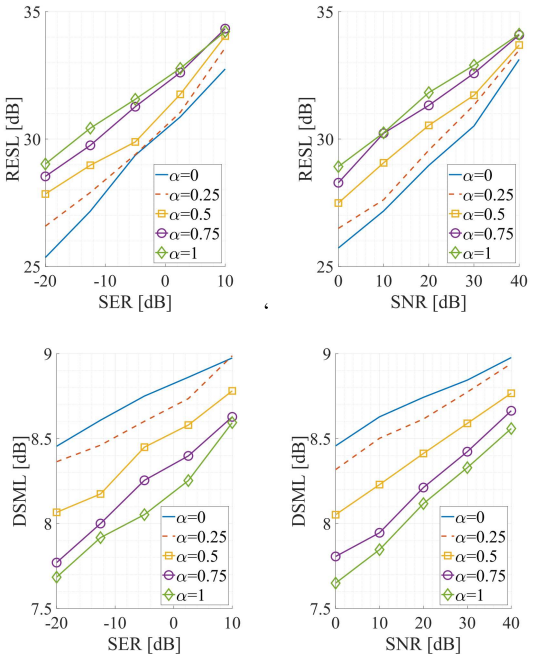


Figure 5: DSML-RESL tradeoff for various values of α in different echo and noise levels.

7. CONCLUSION

We introduced two objective metrics to separately assess the desired-speech maintained level (DSML) and the residual-echo suppression level (RESL) during double-talk. The performances of these metrics are evaluated using a deep learning-based RES system with a tunable design parameter α, with 280 h of real and synthetic recordings. We showed that the DSML and RESL correlate well with human perception compared to the popular SDR metric, which may suggest they are more suitable for speech quality evaluation. Also, we empirically learned the relation between tuning α and the resulting DSML-RESL tradeoff and offered a practical design scheme that benefits dynamic user preferences. Future work will analyze the DNSMOS as an appropriate evaluation for RES subjective quality in double-talk, and explore the DSML-RESL tradeoff to yield a practical design scheme for optimal speech quality.

8. REFERENCES

[1] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *et al.*, "Advances in network and acoustic echo cancellation," 2001.

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

[4] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, 2003.

[5] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[6] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[7] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1827–1831, 2019.

[8] A. Ivry, I. Cohen, and B. Berdugo, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. Interspeech*. IEEE, Sept. 2021.

[9] A. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. WASPAA*. IEEE, 1995, pp. 103–106.

[10] M. I. Mossi, N. W. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *Proc. ICASSP*. IEEE, 2010, pp. 313–316.

[11] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *preprint arXiv:1909.08050*, 2019.

[12] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *preprint arXiv:2010.15258*, 2020.

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

[14] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *Proc. ICASSP*. IEEE, 2018, pp. 231–235.

[15] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Online estimation of reverberation parameters for late residual echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 77–91, 2019.

[16] L. Pfeifenberger and F. Pernkopf, "Nonlinear residual echo suppression using a recurrent neural network," in *Proc. Interspeech*, 2020.

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

[18] B. Fang, "A robust residual echo suppression algorithm even during double talk," in *Proc. 3rd International Conference on Information Communication and Signal Processing (ICICSP)*. IEEE, 2020, pp. 6–9.

[19] ——, "An integrated system of adaptive echo cancellation and residual echo suppression," in *Proc. International Conference on Computer Communication and Information Systems*, 2020, pp. 19–23.

[20] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *Proc. ICASSP*, June 2021.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

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

[23] R. Cutler, A. Saabas, T. Parnamaa, M. Loida, S. Sootla, H. Gamper, *et al.*, "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*. IEEE, Sept. 2021.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. LDC93S1, 1993.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[26] *ITU-T Rec. G.168: Digital network echo cancellers*, ITU-T Std., Feb. 2012.

[27] *ITU-T Rec. P.862: 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-T Std., Feb. 2001.

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

[29] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[30] T. D. Gauthier, "Detecting trends using Spearman's rank correlation coefficient," *Environmental forensics*, vol. 2, no. 4, pp. 359–362, 2001.

[31] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, H. Gamper, S. Braun, *et al.*, "ICASSP 2021 acoustic echo cancellation challenge: Datasets and testing framework," *preprint arXiv:2009.04972*, 2020.