Interspeech 2022
18-22 September 2022, Incheon, Korea

# Objective Metrics to Evaluate Residual-Echo Suppression During Double-Talk in the Stereophonic Case

*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
sivry@campus.technion.ac.il, icohen@ee.technion.ac.il, bbaruch@technion.ac.il

## Abstract

Speech quality, as evaluated by humans, is most accurately assessed by subjective human ratings. The objective acoustic echo cancellation mean opinion score (AECMOS) metric was recently introduced and achieved high accuracy in predicting human perception during double-talk. Residual-echo suppression (RES) systems, however, employ the signal-to-distortion ratio (SDR) metric to quantify speech-quality in double-talk. In this study, we focus on stereophonic acoustic echo cancellation, and show that the stereo SDR (SSDR) poorly correlates with subjective human ratings according to the AECMOS, since the SSDR is influenced by both distortion of desired speech and presence of residual-echo. We introduce a pair of objective metrics that distinctly assess the stereo desired-speech maintained level (SDSML) and stereo residual-echo suppression level (SRESL) during double-talk. By employing a tunable RES system based on deep learning and using 100 hours of real and simulated recordings, the SDSML and SRESL metrics show high correlation with the AECMOS across various setups. We also investigate into how the design parameter governs the SDSML-SRESL tradeoff, and harness this relation to allow optimal performance for frequently-changing user demands in practical cases.

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

## 1. Introduction

A conversation between a pair of speakers, based in near-end and far-end points, is common in hands-free communication. The desired-speech captured by the near-end microphone can be interrupted by echo, which is created by a loudspeaker that emits nonlinearly-distorted version of the far-end signal that reverberates in the room, and by additional noises [1]. An acoustic coupling between the loudspeaker and the microphone potentially occurs due to this echo presence, which impairs the quality of acoustic information transmitted to the far-end [2]. In stereophonic acoustic echo cancellation (SAEC), the echo paths between a pair of loudspeakers and a pair of microphones are modeled by adaptive filtering. The echo paths are converted into acoustic-echo approximations that are subtracted from the microphones [3, 4]. Double-talk segments are most challenging, since the echoes overlap with desired speech. Various studies tried to cope with it by preserving the speech and removing the echoes [5–12]. In practice, however, echo paths are not estimated accurately, e.g., when the adaptive filter has not yet converged [1]. Therefore, a residual-echo suppression (RES) system must succeed the SAEC system to eliminate the echoes.

Subjective human evaluation is currently the most accurate assessment of human perception for speech quality [13,14]. Recently, an objective metric called the acoustic echo cancellation mean opinion score (AECMOS) was introduced. In double-talk specifically, the AECMOS has obtained impressive accuracy in estimating human ratings [15]. In contrast, RES systems conventionally use the signal-to-distortion ratio (SDR) metric [16] to assess speech quality in double-talk, e.g., in [17–24]. It will be empirically shown that the stereo SDR (SSDR) is by definition influenced by both distortion of stereo speech and presence of stereo residual-echo. Thus, for the task of RES in the stereophonic case, the SSDR is not an adequate indicator of neither the human evaluation for quality of speech nor of the AECMOS.

To combat it, we introduce a pair of objective metrics to distinctly assess the stereo desired-speech maintained level (SDSML) and the stereo residual-echo suppression level (SRESL) in double-talk. We first consider an RES system that acts as a time-dependent gain, with a pair of input and output channels. To calculate the SDSML, this gain is projected into the stereo desired-speech and the result is substituted inside the SSDR expression. The SRESL requires an estimate of the noisy stereo residual-echo, achieved by subtracting the stereo desired-speech from the double-talk frame. The ratio between this estimate without and with the gain applied to it generates the SRESL. The SDSML and SRESL metrics are evaluated with an RES system, based on deep learning, which incorporates a tunable design parameter. This study employs 100 h of recordings that comprise of real signals and of simulations in various acoustic setups, with a range of echo and noise levels. Results reveal the AECMOS is well correlated with the SDSML and SRESL with high generalization to various scenarios. An additional empirical study investigates how the design parameter affects the tradeoff between the SDSML and SRESL. We then show how varying the design parameter during training can benefit interchangeable user demands of the RES system, which often occur in real-life. This study extends a recent work by the authors, which address the monophonic AEC case [25].

## 2. Problem Formulation

The RES scenario in the stereophonic case is detailed in Figure 1. Here, bold letters notate vectors and matrices, and normal letters notate scalars. The left and right near-end microphones $m_{\mathrm{L}}(n)$ and $m_{\mathrm{R}}(n)$ at time index $n$ are respectively:

$$m_{\mathrm{L}}(n) = y_{\mathrm{L}}(n) + s_{\mathrm{L}}(n) + w_{\mathrm{L}}(n), \tag{1}$$

$$m_{\mathrm{R}}(n) = y_{\mathrm{R}}(n) + s_{\mathrm{R}}(n) + w_{\mathrm{R}}(n), \tag{2}$$

where $s_{\mathrm{L}}(n)$ and $s_{\mathrm{R}}(n)$ are the near-end speech signals, $w_{\mathrm{L}}(n)$ and $w_{\mathrm{R}}(n)$ represent environmental and system noises, and $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ are the nonlinear reverberant echo signals, as correspondingly captured by the left and right microphones:

$$y_{\mathrm{L}}(n) = \mathbf{h}_{\mathrm{LL}}^{T}(n)\,\mathbf{x}_{\mathrm{NL,L}}(n) + \mathbf{h}_{\mathrm{RL}}^{T}(n)\,\mathbf{x}_{\mathrm{NL,R}}(n), \tag{3}$$

$$y_{\mathrm{R}}(n) = \mathbf{h}_{\mathrm{LR}}^{T}(n)\,\mathbf{x}_{\mathrm{NL,L}}(n) + \mathbf{h}_{\mathrm{RR}}^{T}(n)\,\mathbf{x}_{\mathrm{NL,R}}(n). \tag{4}$$

10.21437/Interspeech.2022-673

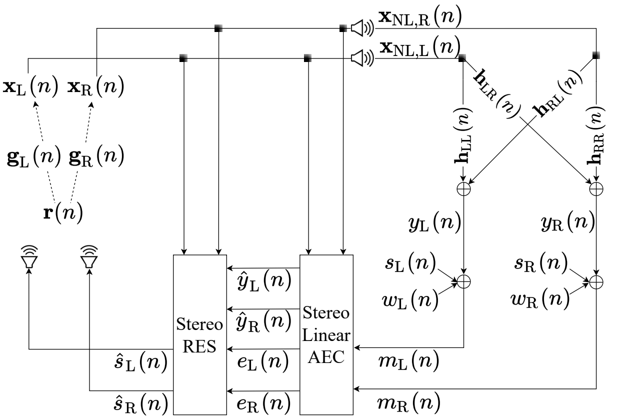

Figure 1: *RES scenario in the stereophonic case.*

Here, $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ respectively denote the $L$ last samples of the left and right far-end signals, $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$, after nonlinear distortions by nonideal hardware [26]:

$$\mathbf{x}_{\text{NL,L}}(n) = \left[x_{\text{NL,L}}(n), \ldots, x_{\text{NL,L}}(n-L+1)\right]^T, \quad (5)$$

$$\mathbf{x}_{\text{NL,R}}(n) = \left[x_{\text{NL,R}}(n), \ldots, x_{\text{NL,R}}(n-L+1)\right]^T, \quad (6)$$

and each of the column vectors $\mathbf{h}_{\text{LL}}(n)$, $\mathbf{h}_{\text{RL}}(n)$, $\mathbf{h}_{\text{LR}}(n)$, $\mathbf{h}_{\text{RR}}(n)$ has $L$ samples and represents a room impulse response (RIR) from the loudspeakers to the microphones. Preliminary, linear echo is reduced by employing the system in [27]. This system receives $m_{\text{L}}(n)$ and $m_{\text{R}}(n)$ as inputs, and $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$ as reference channels, and generates two pairs of signals: a pair of echo estimates $\hat{y}_{\text{L}}(n)$ and $\hat{y}_{\text{R}}(n)$, and a pair of near-end speech signal estimates $e_{\text{L}}(n)$ and $e_{\text{R}}(n)$, given by

$$\begin{aligned} e_{\text{L}}(n) &= m_{\text{L}}(n) - \hat{y}_{\text{L}}(n) && (7)\\ &= \left(y_{\text{L}}(n) - \hat{y}_{\text{L}}(n)\right) + s_{\text{L}}(n) + w_{\text{L}}(n), \\ e_{\text{R}}(n) &= m_{\text{R}}(n) - \hat{y}_{\text{R}}(n) && (8)\\ &= \left(y_{\text{R}}(n) - \hat{y}_{\text{R}}(n)\right) + s_{\text{R}}(n) + w_{\text{R}}(n). \end{aligned}$$

The RES system aims to suppress the residual echoes, i.e., both $y_{\text{L}}(n) - \hat{y}_{\text{L}}(n)$ and $y_{\text{R}}(n) - \hat{y}_{\text{R}}(n)$, without distorting the desired-speech signals, i.e., $s_{\text{L}}(n)$ and $s_{\text{R}}(n)$.

## 3. The SDSML and SRESL Metrics

The SDSML and SRESL are developed by assuming a two-input and two-output RES system that acts as a time-varying gain matrix. The gain matrix in double-talk periods is given by

$$\mathbf{g}(n) = 0.5 \begin{bmatrix} \hat{s}_{\text{L}}(n)/e_{\text{L}}(n) & \hat{s}_{\text{L}}(n)/e_{\text{R}}(n) \\ \hat{s}_{\text{R}}(n)/e_{\text{L}}(n) & \hat{s}_{\text{R}}(n)/e_{\text{R}}(n) \end{bmatrix}, \quad (9)$$

where in double-talk $e_{\text{L}}(n) \neq 0$ and $e_{\text{R}}(n) \neq 0$. Before introducing the SDSML and SRESL definitions, we inspect the shortcomings of the SSDR. Extending the traditional SDR definition [16] to the stereophonic case, it follows that

$$\begin{aligned} \text{SSDR} &= 10\log_{10} \left.\frac{\|\mathbf{s}(n)\|_2^2}{\|\mathbf{s}(n) - \hat{\mathbf{s}}(n)\|_2^2}\right|_{\text{Double-talk}} \\ &= 10\log_{10} \left.\frac{\|\mathbf{s}(n)\|_2^2}{\|\mathbf{s}(n) - \mathbf{g}(n)\mathbf{e}(n)\|_2^2}\right|_{\text{Double-talk}}, \end{aligned} \quad (10)$$

where

$$\mathbf{s}(n) = \begin{bmatrix} s_{\text{L}}(n) \\ s_{\text{R}}(n) \end{bmatrix}, \ \hat{\mathbf{s}}(n) = \begin{bmatrix} \hat{s}_{\text{L}}(n) \\ \hat{s}_{\text{R}}(n) \end{bmatrix}, \ \mathbf{e}(n) = \begin{bmatrix} e_{\text{L}}(n) \\ e_{\text{R}}(n) \end{bmatrix}. \quad (11)$$

Both stereo desired-speech distortion and stereo residual-echo presence influence the SSDR value. Since the SSDR employs the term $\mathbf{g}(n)\mathbf{e}(n)$, a scenario of distortion-free speech and echo and a scenario of distorted speech without echo may produce an identical SSDR value. These scenarios, however, are perceived differently by humans and present different AEC-MOS values. It will be empirically shown that the SSDR and subjective human perception are poorly matched according to the AECMOS. Reliable evaluation of RES systems during double-talk can be achieved by separating the quantification of speech distortion from one of residual-echo suppression. Such distinction is not provided by the AECMOS metric. Thus, a pair of objective metrics is introduced by separately employing $\mathbf{g}(n)$ to the stereo desired-speech and to the noisy stereo residual-echo estimate.

The SDSML definition is analogous to the SSDR, except that $\mathbf{g}(n)$ is projected to the stereo desired-speech $\mathbf{s}(n)$ solely:

$$\text{SDSML} = 10\log_{10} \left.\frac{\|\tilde{\mathbf{s}}(n)\|_2^2}{\|\tilde{\mathbf{s}}(n) - \mathbf{g}(n)\mathbf{s}(n)\|_2^2}\right|_{\text{Double-talk}}. \quad (12)$$

Next, the noisy stereo residual-echo is evaluated as $\mathbf{r}(n) = \mathbf{e}(n) - \mathbf{s}(n)$, and the SRESL is calculated by:

$$\text{SRESL} = 10\log_{10} \left.\frac{\|\mathbf{r}(n)\|_2^2}{\|\mathbf{g}(n)\mathbf{r}(n)\|_2^2}\right|_{\text{Double-talk}}. \quad (13)$$

It is noted that a constant attenuation may occur by the RES system, which deviates the SDSML from its real value. The SDSML must be unvaried by this attenuation, so it is being restored as shown in [25]. Expressly, $\tilde{\mathbf{s}}(n) = \tilde{g}(n)\mathbf{s}(n)$, where:

$$\tilde{g}(n) = \frac{\langle \mathbf{g}(n)\mathbf{s}(n), \mathbf{s}(n) \rangle}{\|\mathbf{s}(n)\|_2^2}. \quad (14)$$

## 4. A Tunable Stereophonic RES System

An RES system based on deep learning, inspired by [26], is employed to assess the SDSML and SRESL metrics. It contains six input channels, and two output channels and operates in the waveform domain. The proposed architecture is comprised of blocks of nonlinear models (NLMs). Each NLM comprises 3 gated recurrent units (GRUs) that contain 16 cells each [28] and dropout [29] in the recurrent layers, an FCNN with a two-neuron output, and a piecewise linear unit (PLU) activation function with trainable parameters [30] that is applied to each output neuron. The architecture is modeled by 3 consecutive NLMs. The first NLM receives the outputs of the linear SAEC system, i.e. $\hat{y}_{\text{L}}(n)$, $\hat{y}_{\text{R}}(n)$, $e_{\text{L}}(n)$, $e_{\text{R}}(n)$, and the two reference channels $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$, and emits two output channels. The two succeeding NLMs are fed with four entrances each; a pair of output channels of the previous NLM, and the two reference channels. The last NLM produces the speech estimates $\hat{s}_{\text{L}}(n)$ and $\hat{s}_{\text{R}}(n)$. A tunable design parameter $0 \leq \alpha \leq 1$, originally introduced in [31], controls an intrinsic tradeoff that occurs inside a customized loss function $J(\alpha)$:

$$\begin{aligned} J(\alpha) &= \alpha\,\text{SDSML}^{-1} + (1-\alpha)\,\text{SRESL}^{-1} \\ &= \alpha \left[\frac{\|\tilde{\mathbf{s}}(n)\|_2^2}{\|\tilde{\mathbf{s}}(n) - \mathbf{g}(n)\mathbf{s}(n)\|_2^2}\right]^{-1} \\ &\quad + (1-\alpha) \left[\frac{\|\mathbf{r}(n)\|_2^2}{\|\mathbf{g}(n)\mathbf{r}(n)\|_2^2}\right]^{-1}, \end{aligned} \quad (15)$$

where during double-talk $\tilde{\mathbf{s}}(n), \mathbf{r}(n) \neq \mathbf{0}$ and $\mathbf{0}$ is a vector of zeros. The parameter $\alpha$ compromises between the SDSML and SRESL values in the training stage while $J(\alpha)$ is minimized. As a result, the stereo desired-speech distortion and stereo residual-echo suppression levels that the system permits can be adjusted dynamically. For instance, setting $\alpha = 1$ forces the stereo desired-speech prediction to coincide with its ground truth. Shifting to $\alpha = 0$, however, focuses on suppressing the stereo residual-echo, but causes a more substantial stereo desired-speech distortion. Tuning $\alpha$, i.e., tuning the SDSML-SRESL tradeoff, can be done dynamically during training.

This RES system contains 23 thousand parameters that consume 550 million floating-point operations per second (Mflops) and 65 KB of memory. Its embedding on hands-free platforms is thus feasible, e.g., by considering the NDP120 neural processor by Syntiant$^{\text{TM}}$ [32]. The preceding linear AEC system employs the sign-error normalized least mean squares (SNLMS) adaptive filter in the sub-band domain [27].

# 5. Experimental Setup

## 5.1. Database

This study makes use of the AEC challenge database [33] that is sampled at 16 kHz and incorporates English double-talk segments both with and without echo-paths change. In scenarios of no echo-paths change, the near-end setup does not include movements. In scenarios of echo-paths change, however, the recording involves movement in the near-end, either by the speaker or the device. This database contains 75 h of real clean and noisy recordings and additional 25 h of synthetic data, which are assigned to the original far-end source signal $\mathbf{r}(n)$ and to the near-end speech and noise signals, where $s_{\text{L}}(n) = s_{\text{R}}(n)$ and $w_{\text{L}}(n) = w_{\text{R}}(n)$ in this study. To produce the far-end signals $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$, $\mathbf{r}(n)$ is randomly propagated via one of 4500 pairs of RIRs that generate $\mathbf{g}_{\text{L}}(n)$ and $\mathbf{g}_{\text{R}}(n)$, i.e., the acoustic paths between $\mathbf{r}(n)$ and the left and right far-end microphones, respectively. To account for realistic acoustic environments, $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$ randomly undergo one of 4500 artificial nonlinearities that confine with practical characteristics of power amplifiers and loudspeakers in modern hands-free devices [26]. Each pair of nonlinearly-distorted far-end signals $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ is randomly propagated via one of 4500 foursomes of near-end RIRs. All RIRs are generated using the Image Method [34] with $L$ coefficients and reverberation times RT$_{60}$, where RT$_{60} \sim U[0.2, 0.5]$ seconds. The near-end stereo-speech-to-echo ratio (SSER) and stereo-speech-to-noise ratio (SSNR) levels were distributed on $[-10, 10]$ dB and $[0, 40]$ dB, respectively, and are defined as SSER=$10\log_{10}\left[\|\mathbf{s}(n)\|_2^2 / \|\mathbf{y}(n)\|_2^2\right]$ and SSNR=$10\log_{10}\left[\|\mathbf{s}(n)\|_2^2 / \|\mathbf{w}(n)\|_2^2\right]$ in dB, where both $\mathbf{y}(n)$ and $\mathbf{w}(n)$ follow the notations in eq. (11) and both ratios are calculated with 50% overlapping time frames of 5 seconds.

## 5.2. Data Processing, Training, and Testing

The entire database is divided into 80 h of training, 10 h of validation, and 10 h of test sets in a random manner. Bias is averted by following conventions that balance all sets [31]. Since real-life scenarios often involve an abrupt change in the echo paths, we simulate these to reoccur every $t$ seconds, where $t \sim U[4, 10]$, and set $L = 2400$. The NN is fed with 50% overlapping time frames of 20 ms and is trained with a learning rate of $10^{-4}$ that decays by $10^{-6}$ every 5 epochs, mini-batch size of 60 ms, and 40 epochs, using Adam optimizer [35] and back-propagation through time. On average, training the RES system lasted 25 minutes for every 10 h of data and inference took 8 ms per batch on an Intel Core i7-8700K CPU @ 3.7 GHz with two GPUs of type Nvidia GeForce RTX 2080 Ti.

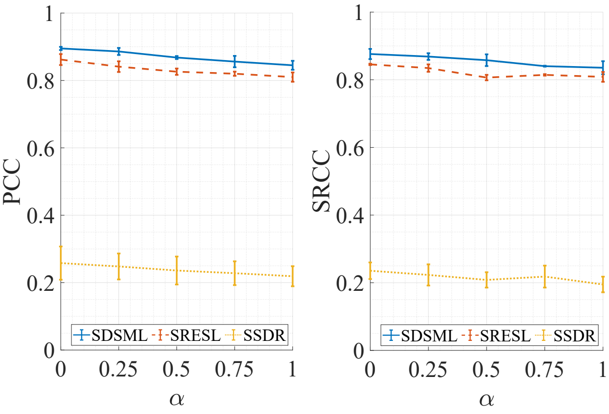

Figure 2: *Correlation of the AECMOS with the SDSML, SRESL, and SSDR metrics.*

## 5.3. Additional Performance Metrics

Performance is also evaluated with the SSDR, which is influenced by both echo presence and distortion of speech, and with the perceptual evaluation of speech quality (PESQ) [36] between $\mathbf{s}(n)$ and $\hat{\mathbf{s}}(n)$. The AECMOS is also reported, and is calculated using the API provided by Microsoft as the average between the AECMOS of $\hat{\mathbf{s}}_{\text{L}}(n)$ and the AECMOS of $\hat{\mathbf{s}}_{\text{R}}(n)$.

# 6. Experimental Results

Results are reported on the test set. In Tables 1 and 2, both mean and standard deviation (std) values are given. In Figure 4, only mean values are shown. Higher mean and lower std values entail better performance for all metrics. The linear filter convergence confines with the description in [27, 37]. We use 50% overlapping time frames of 5 seconds for metrics calculations.

We employ the Pearson correlation coefficient (PCC) [38] and Spearman's rank correlation coefficient (SRCC) [39] to discover how much the SDSML and SRESL correlate with the AECMOS, similarly to [15, 40, 41]. This experiment includes segments both with and without echo-paths change after the linear SAEC system has converged for $\alpha = [0, 0.25, 0.5, 0.75, 1]$, and the results are shown in Figure 2. The SSDR and AECMOS are poorly correlated, as pointed out by the PCC and SRCC mean values that fall below 0.26 for all $\alpha$. However, with average PCC and SRCC values between 0.8 and 0.89 for all $\alpha$, the proposed SDSML and SRESL metrics highly coordinate with the AECMOS. Observing std values, the SDSML and SRESL show more consistent correlations across $\alpha$ than the SSDR. Figure 3 visualizes the AECMOS versus the SDSML, SRESL, and SSDR metrics for random sample values drawn from $\alpha \sim U[0.25, 0.75]$ with no echo-paths change. The low matching between the AECMOS and the SSDR and the high correlation between the AECMOS and the SDSML and SRESL are now verified. The SDSML and SRESL are therefore more indicative to subjective human perception of speech-quality evaluation than the SSDR, according to the AECMOS.

In Tables 1 and 2 performance metrics are evaluated for scenarios without and with echo-paths change, respectively, after convergence with $\alpha = [0, 0.5, 1]$. The trend of the SDSML and SRESL is consistent with the one of the SSDR, all of which

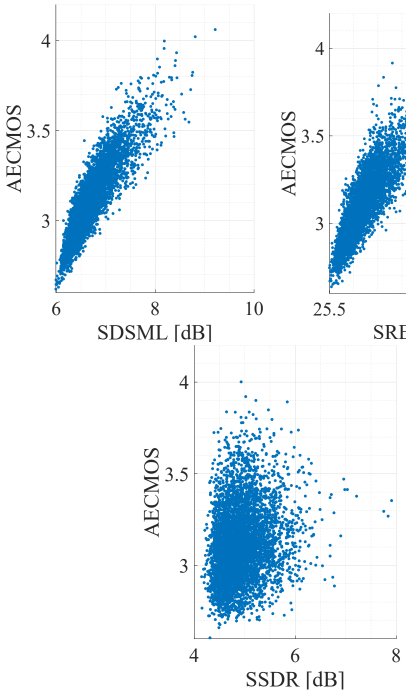

Figure 3: *Scatter plots of the AECMOS versus the SDSML, SRESL, and SSDR metrics.*

deteriorate in the transition from no echo-paths change to echo-paths change periods. This consistency across various test set setups indicates high generalization of the SDSML and SRESL. The average SDSML values are regularly higher than the average SSDR values, as expected. This is directly derived from eq. (10), in which the denominator takes into account both echo and noise. As values of $\alpha$ increase, the average SDSML values increase while the average SRESL values decrease, both with and without echo-paths change, as expected.

We now explore how $\alpha$ governs the tradeoff between the SDSML and SRESL. In Figure 4, results for no echo-paths change periods after convergence are included, for $\alpha = [0, 0.25, 0.5, 0.75, 1]$. Lower $\alpha$ values relate to lower SDSML values because distortion is higher for stereo speech. The SRESL values rise, however, since more suppression is applied to the stereo residual-echo. Empirically, this tradeoff is consistent on average for all $\alpha$. We also explore how practical SSER and SSNR levels influence this tradeoff. As expected, the more acoustic conditions degrade, the more the values of both metrics are impaired. Also, regardless of acoustic conditions, it is maintained that the lower $\alpha$ becomes, the lower the SDSML and the higher SRESL values appear.

Practical user demands of the RES system may vary. Thus, we propose a design scheme that addresses this dynamic need. As an example, let us assume convergence has been achieved and no echo-paths change occurs. This can be verified by following the definitions in [27, 37]. Initially, the user requires an average SRESL higher than 24.5 dB and an average SDSML higher than 7.6 dB. They inspect Figure 4 and select $\alpha = 0.75$. At some point, the user concludes that SSER $= 0$ dB and SSNR $= 20$ dB, e.g., by respectively analyzing double-talk and near-end single-talk periods. Thus, they demand a SDSML no lower than 7.4 dB with a maximal SRESL. Again, the user follows Figure 4 and decides to shift $\alpha = 0.75$ to $\alpha = 0.5$ during training. Indeed, this lowers the average SDSML to 7.45 dB and enhances the average SRESL to over 26.5 dB. Note that the conclusions in this section also hold for the mono AEC case [25].

Table 1: *Performance with no echo-paths change.*

| $\alpha$ | SDSML | SRESL | SSDR |
|---|---|---|---|
| 0 | 6.15±0.6 | 29.6±3.3 | 4.55±0.9 |
| 0.5 | 7.38±0.6 | 26.4±3.4 | 5.52±0.8 |
| 1 | 8.13±0.5 | 23.1±3.8 | 6.73±0.7 |

Table 2: *Performance with echo-paths change.*

| $\alpha$ | SDSML | SRESL | SSDR |
|---|---|---|---|
| 0 | 5.41±1.1 | 24.3±3.5 | 3.61±1.4 |
| 0.5 | 6.29±1.0 | 21.7±4.0 | 4.60±1.2 |
| 1 | 7.01±0.8 | 18.9±4.9 | 5.54±1.0 |

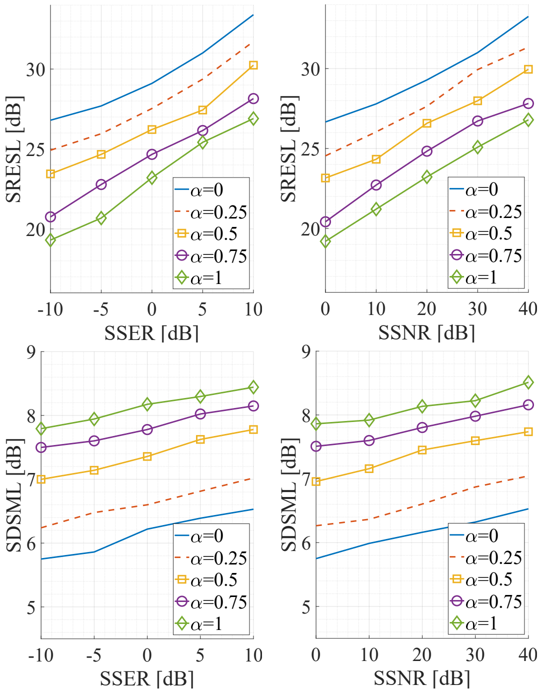

Figure 4: *SDSML-SRESL tradeoff for various values of $\alpha$.*

## 7. Conclusions

We focused on the task of RES in the stereophonic case during double-talk. We first showed that the widely-used SSDR metric poorly correlates with human speech-quality ratings. We then proposed a pair of objective measures that distinct between desired-speech distortion and residual-echo suppression during double-talk. By considering a deep RES system with a tunable parameter $\alpha$, we showed that the SDSML and SRESL correlate well with the AECMOS metric, which may render they are more appropriate to assess quality of speech. Also, by tuning $\alpha$ during training, we offered a practical design scheme that allows flexible adjustment of the RES system to a specific SDSML-SRESL tradeoff. A sequential study will focus on enhancing subjective experience for RES systems during double-talk periods by optimizing the AECMOS through tuning of $\alpha$.

# 8. References

[1] 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.

[2] A. Gilloire and M. Vetterli, "Adaptive filtering in sub-bands with critical sampling: analysis, experiments, and application to Acoust. echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, 1992.

[3] 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.

[4] 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 Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 156–165, 1998.

[5] C. Stanciu, J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "A widely linear model for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 93, no. 2, pp. 511–516, 2013.

[6] C. Paleologu, J. Benesty, and S. Ciochină, "Widely linear general kalman filter for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 94, pp. 570–575, 2014.

[7] S. Cecchi, L. Romoli, P. Peretti, and F. Piazza, "Low-complexity implementation of a real-time decorrelation algorithm for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 92, no. 11, pp. 2668–2675, 2012.

[8] A. Kar and M. Swamy, "Tap-length optimization of adaptive filters used in stereophonic acoustic echo cancellation," *Signal Process.*, vol. 131, pp. 422–433, 2017.

[9] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 686–696, 2001.

[10] S. Wu, X. Qiu, and M. Wu, "Stereo acoustic echo cancellation employing frequency-domain preprocessing and adaptive filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 614–623, 2010.

[11] L. Romoli, S. Cecchi, L. Palestini, P. Peretti, and F. Piazza, "A novel approach to channel decorrelation for stereo acoustic echo cancellation based on missing fundamental theory," in *Proc. Int. Conf. Acoust., Speech and Signal Process.* IEEE, 2010, pp. 329–332.

[12] A. Gilloire and V. Turbin, "Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers," in *Proc. ICASSP*, vol. 6. IEEE, 1998, pp. 3681–3684.

[13] 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.

[14] R. Cutler, B. Nadari, M. Loide, S. Sootla, and A. Saabas, "Crowdsourcing approach for subjective evaluation of echo impairment," in *Proc. ICASSP*. IEEE, 2021, pp. 406–410.

[15] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "AECMOS: A speech quality assessment metric for echo impairment," *preprint arXiv:2110.03010*, 2021.

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

[17] 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.

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

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

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

[21] B. Fang, "A robust residual echo suppression algorithm even during double talk," in *Proc. Int. Conf. Info. Comm. Signal Process. (ICICSP)*. IEEE, 2020, pp. 6–9.

[22] ——, "An integrated system of adaptive echo cancellation and residual echo suppression," in *Proc. Int. Conf. Comput. Comm. Info. Sys.*, 2020, pp. 19–23.

[23] T. S. Wada and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 175–189, 2011.

[24] E. Kim, J.-J. Jeon, and H. Seo, "U-convolution based residual echo suppression with multiple encoders," in *Proc. ICASSP*. IEEE, 2021, pp. 925–929.

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

[26] ——, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. Interspeech*, 2021, pp. 4773–4777.

[27] ——, "Deep adaptation control for stereophonic acoustic echo cancellation," in *Proc. Interspeech*, 2022, submitted.

[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *preprint arXiv:1412.3555*, 2014.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] A. Nicolae, "PLU: The piecewise linear unit activation function," *preprint arXiv:1809.09534*, 2018.

[31] 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.

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

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

[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[36] *ITU-T Rec. P.862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Std., Oct. 2017.

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

[38] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Process.* Springer, 2009, pp. 1–4.

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

[40] 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.

[41] 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.