ACOUSTIC ECHO REDUCTION ROBUST AGAINST ECHO-PATH CHANGE WITH INSTANT ECHO-POWER-LEVEL ADJUSTMENT

[1]*Masahiro Fukui,* [1]*Suehiro Shimauchi,* [2]*Yusuke Hioka,* [1]*Hitoshi Ohmuro, and* [3]*Yoichi Haneda*

[1]NTT Media Intelligence Laboratories, NTT Corporation,
3-9-11, Midori-cho, Musashino-shi, Tokyo 180-8585, Japan
[2]Department of Electrical and Computer Engineering, University of Canterbury,
20 Kirkwood Ave, Ilam, Christchurch 8041, New Zealand
[3]Faculty of Informatics and Engineering, The University of Electro-Communications,
1-5-1, Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
phone: +81 422 59 2113, fax: +81 422 60 7811, email: fukuimas@ieee.org

ABSTRACT

This paper proposes a new method of residual echo reduction (ER) to track abrupt increases in the residual echo. The method assumes that the spectral structure of the residual echo is maintained even if the residual echo abruptly increases; then, estimates only of the echo power level can be made in a short observation time. The proposed method instantaneously obtains the echo power level by focusing on all frequency-spectral components, and recalculates the echo component using bothe the obtained power level and the spectral structure estimated with the conventional method. The outstanding performance of the proposed method is demonstrated through objective and subjective experiments. The results revealed that the degradation in speech quality was mitigated even if the echo power level increased during the double-talk periods.

Index Terms— Acoustic echo canceller, residual echo, STSA-based echo reduction, echo-path change

1. INTRODUCTION

An acoustic echo canceller (AEC) is used in hands-free telecommunication systems to eliminate unwanted acoustic echo signals that result from acoustic coupling between the loudspeaker and microphone. In most AECs, an adaptive filter (ADF) [1, 2, 3] and echo reduction (ER) [4, 5, 6, 7, 8, 9, 10, 11] are jointly employed in series. Although the ADF generates an echo replica, which is subtracted from the microphone signal, some residual echo still remains in its output (often called an *error signal*). The ER follows the ADF to reduce the residual echo [12, 13]. This process suppresses the residual echo by multiplying echo-reduction gain to each frequency component of the error signal. The echo-reduction gain is calculated from the estimate of the acoustic coupling level (ACL), which is a power frequency response of the transfer function between the loudspeaker and microphone.

There are two representative methods of estimating ACL; the first is straightforward and simple [4, 5], which measures ACL when the near-end speech is detected to be absent; the second is a method of frequency-domain cross correlation (FDCC) [6, 7] using the long-time correlation between received and error signals. A major advantage of the FDCC method is that it has the ability to estimate ACL even during double-talk periods by exploiting the uncorrelatedness between the received and near-end speech signals. Therefore, the FDCC approach is gradually becoming the mainstream for ACL estimation. However, a problem with FDCC is its slow tracking speed because the uncorrelatedness assumption only holds true if the observation period (time period) of the signals is long enough. As a result, FDCC fails to accurately track immediate increases in the residual echo level that occur during the double-talk period or when the echo-path abruptly changes. If the length of a signal utilized to estimate ACL is shortend to obtain a faster tracking speed, the estimation accuracy of ACL will be degraded because the uncorrelatedness assumption is no longer valid.

To resolve this trade-off, we propose a new method of echo reduction with instantaneous adjustment of the echo power level. This method is based on a model that assumes the power spectral structure of the residual echo does not change even if the amount of residual echo abruptly increases. Because the power spectrum of a signal is determined by the spectral structure and the power level, the model allows the proposed method to instantaneously estimate the amount of residual echo by only estimating the power level of the residual echo spectrum. The proposed method calculates the instantaneous echo power level even during the double-talk period by focusing on all frequency samples in the cross-power spectral calculation between the error signal and the estimated residual echo from FDCC, and the resulting power level helps to accurately estimate the amount of residual echo in the error signal. Thus, the proposed method is able to reduce echo after the abrupt increase in the residual echo, while retaining the accuracy of estimating the echo-reduction gain even during double-talk periods.

The remainder of this paper is organized as follows. Section 2 presents an overview of the conventional ER method and its problems. Section 3 explains the proposed echo-power-level adjustment method. Section 4 presents results from demonstrations of the performance of the proposed method through simulations and subjective evaluations, and Section 5 concludes the paper.

2. ECHO REDUCTION AND ACCOMPANYING PROBLEMS

This section explains the echo-reduction (ER) process in a short-time spectral amplitude (STSA) domain [14] using the method of frequency-domain cross correlation (FDCC). The structure of ER is outlined in Fig. 1, and the adaptive filter (ADF) process has been omitted to simplify the explanation. Microphone signal $y(n)$ is expressed as

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

where $d(n)$ is the acoustic echo signal and $s(n)$ near-end speech signal. Here, $d(n)$ is represented by the convolution of echo-path impulse response $h(n)$ and received signal $x(n)$, i.e.,

$$d(n) = h(n) * x(n), \tag{2}$$

where $*$ denotes the convolution. The short-time spectrums of $d(n)$ and $y(n)$ are respectively represented as

$$D_i(\omega) = H_i(\omega)X_i(\omega) \tag{3}$$

and

$$Y_i(\omega) = D_i(\omega) + S_i(\omega), \tag{4}$$

where ω is the discrete frequency index, i is the discrete time-frame index, and $H_i(\omega)$, $X_i(\omega)$, and $S_i(\omega)$ correspond to the short-time spectra of $h(n)$, $x(n)$, and $s(n)$, respectively.

The echo reduction in general form can be expressed by

$$\hat{S}_i(\omega) = G_i(\omega)Y_i(\omega), \tag{5}$$

where $\hat{S}_i(\omega)$ denotes the estimate of $S_i(\omega)$. $G_i(\omega)$ is the echo-reduction gain, which is, for example, calculated according to the Wiener filtering method [15, 16] obtained by

$$G_i(\omega) = \frac{|Y_i(\omega)|^2 - |\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2}, \tag{6}$$

where $|\hat{D}_i(\omega)|^2$ is the estimate of echo power spectrum $|D_i(\omega)|^2$. The obtained estimate $\hat{S}_i(\omega)$ is transformed into time domain signal $\hat{s}(n)$, which is the send signal an inverse fast Fourier transform (IFFT).

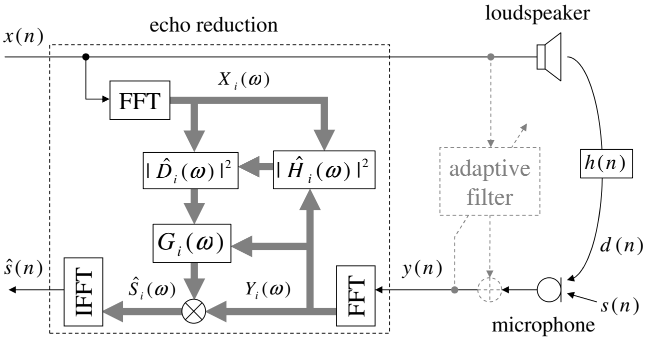

Fig. 1. Structure of echo reduction process.

Echo power spectrum $|D_i(\omega)|^2$ is estimated as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i(\omega)|^2|X_i(\omega)|^2, \tag{7}$$

where $|\hat{H}_i(\omega)|^2$ denotes the estimate of the acoustic coupling level (ACL) $|H_i(\omega)|^2$, which is the power spectrum of $h(n)$. The ACL estimate is given [6] by

$$|\hat{H}_i(\omega)|^2 = \left|\frac{\langle \mathbf{X}_i(\omega), \mathbf{Y}_i(\omega)\rangle}{\|\mathbf{X}_i(\omega)\|^2}\right|^2, \tag{8}$$

where $\langle , \rangle$ and $\|\cdot\|$ are an inner product and a norm, respectively. Boldface denotes a time-sequence vector of a short-time spectrum: $\mathbf{P}_i(\omega) = [P_i(\omega), \cdots, P_{i-L+1}(\omega)]^T$. L is the number of frames, meaning the observation time parameter, and T represents the transposition. The ACL estimate $|\hat{H}_i(\omega)|^2$ is close to zero if the ADF converged properly.

The estimate in (8) is obtained from the following relationship:

$$\begin{aligned} |\hat{H}_i(\omega)|^2 &= \left|\frac{\langle \mathbf{X}_i(\omega), \mathbf{D}_i(\omega)\rangle + \langle \mathbf{X}_i(\omega), \mathbf{S}_i(\omega)\rangle}{\|\mathbf{X}_i(\omega)\|^2}\right|^2, & (9)\\ &\simeq \left|\frac{\langle \mathbf{X}_i(\omega), \mathbf{D}_i(\omega)\rangle}{\|\mathbf{X}_i(\omega)\|^2}\right|^2, & (10)\\ &\simeq \left|\frac{H_i(\omega)\,\langle \mathbf{X}_i(\omega), \mathbf{X}_i(\omega)\rangle}{\|\mathbf{X}_i(\omega)\|^2}\right|^2, & (11)\\ &= |H_i(\omega)|^2. & (12)\end{aligned}$$

However, the approximation from (9) to (10) holds only when the time-sequence vectors of the received and near-end speeches are uncorrelated. Therefore, the conventional method needs to take a large L for the time period because the statistical properties of data are used for the calculation. However, the derivation from (10) to (11) holds only if the echo path does not vary during the past L frames. Thus, the tracking speed of ACL estimation is slow when L is large, and as a result, the conventional method suffers from the tradeoff between tracking speed and accuracy for ACL estimation.

3. PROPOSED METHOD

3.1. Strategy to improve echo reduction

This section proposes a method of adjusting the echo power level to improve the conventional tracking speed while maintaining accuracy by using the large L. Echo-reduction gain in the proposed method is compensated for by introducing a new adjustable parameter of echo-power level $\hat{C}_i$, as follows:

$$G_i^P(\omega) = \frac{|Y_i(\omega)|^2 - \hat{C}_i|\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2}. \tag{13}$$

$\hat{C}_i$ is adaptively estimated and quickly tracks abrupt changes in the echo level (i.e. a overestimation parameter as introduced for noise suppression technique in [17]). To instantaneously estimate the adjustable parameter, $\hat{C}_i$ is derived as

$$\hat{C}_i = \frac{\left\langle \overline{\mathbf{Y}}_i, \overline{\hat{\mathbf{D}}}_i \right\rangle}{\|\overline{\hat{\mathbf{D}}}_i\|^2}, \tag{14}$$

where the boldface with the overline denotes a frequency-sequence vector of a short-time power spectrum at frame index i: $\overline{\mathbf{P}}_i = [|P_i(0)|^2, \cdots, |P_i(\omega)|^2, \cdots, |P_i(N-1)|^2]^T$. N is the number of all frequency bins. In (14), the average of the frequency elements is used instead of the time average. $\hat{C}_i$ is not dependent on the frequency because the proposed method assumes that the spectral structure of the echo remains unchanged, although the echo abruptly increases. Therefore, our method only takes into consideration the instantaneous calculation of the power level of the echo.

If the power spectra of the estimated echo and near-end speech are uncorrelated and the spectral structures between $D_i(\omega)$ and $\hat{D}_i(\omega)$ are similar, $\hat{C}_i$ can correctly adjust the estimate of echo power spectrum:

$$\hat{C}_i\|\overline{\hat{\mathbf{D}}}_i\| \simeq \frac{\left\langle \overline{\mathbf{D}}_i, \overline{\hat{\mathbf{D}}}_i \right\rangle + \left\langle \overline{\mathbf{S}}_i, \overline{\hat{\mathbf{D}}}_i \right\rangle}{\|\overline{\hat{\mathbf{D}}}_i\|^2}\|\overline{\hat{\mathbf{D}}}_i\|, \tag{15}$$

$$\simeq \frac{\left\langle \overline{\mathbf{D}}_i, \overline{\hat{\mathbf{D}}}_i \right\rangle}{\|\overline{\hat{\mathbf{D}}}_i\|^2}\|\overline{\hat{\mathbf{D}}}_i\|, \tag{16}$$

$$\simeq \frac{\|\overline{\mathbf{D}}_i\|}{\|\overline{\hat{\mathbf{D}}}_i\|}\|\overline{\hat{\mathbf{D}}}_i\|, \tag{17}$$

$$= \|\overline{\mathbf{D}}_i\|. \tag{18}$$

The above approximate expressions are equivalent to ignoring the effects of the cross term between $\hat{D}_i(\omega)$ and $S_i(\omega)$, and so they sometimes might lead to some oversubtraction. However, this method can expect to improve the ER performance right after the abrupt increase in the echo.

Equation (14) also corresponds to a least squares solution to the following equation:

$$\|\overline{\mathbf{Y}}_i - \hat{C}_i\overline{\hat{\mathbf{D}}}_i\|^2 \to 0. \tag{19}$$

Table 1. Experimental conditions

Sampling rate	16 kHz
Frame length	256 samples
Frame shift	128 samples
FFT points	256 samples
Reverberation time	300 ms

The adjustable parameter is generally calculated using time and frequency spectral domains for the cross-spectral calculation in order to obtain more statistical data as

$$\hat{C}_i' = \frac{\sum_{m=0}^{M-1}\sum_{\omega=0}^{N-1}|Y_{i-m}(\omega)|^2|\hat{D}_{i-m}(\omega)|^2}{\sum_{m=0}^{M-1}\sum_{\omega=0}^{N-1}|\hat{D}_{i-m}(\omega)|^4}, \tag{20}$$

where M is the number of frames, with $M \ll L$.

3.2. Practical calculation method

This section presents a more practical method of calculating the adjustable parameter to obtain better performance. In practice, the difference between $|D_i(\omega)|^2$ and $|\hat{D}_i(\omega)|^2$ causes problems with accuracy of power-level estimation. The estimation-error component is regarded as the near-end speech component, and as a result, the echo-power-level estimate will be smaller than the actual level. Consequently, the proposed method floors the adjustable parameter of the echo-power level with the underfloor threshold as

$$\hat{C}_i^P = \max\left[\hat{C}_i', \mathrm{Th}\right], \tag{21}$$

where $\max[\cdot]$ is the maximum value selection, and Th is the constant value to determine the minimum value of the adjustable parameter, where $\mathrm{Th} = 1$.

4. EVALUATION

The performance of our new method was evaluated using both simulation and subjective listening tests. The proposed and conventional methods were used to calculate the echo-reduction gain using (13) and (6), respectively. The conventional method is ER with FDCC. The ADF process was skipped in these evaluations in order to evaluate the element performance of ER process. Table 1 lists the experimental conditions.

4.1. Simulation results

The received and near-end speech signals are shown in Figs. 2 (a) and (b), respectively. Periods A and B are received and send single-talk situations, respectively. The double-talk situation occurs during periods C and D. The echo path was changed at 12 seconds into the different impulse response,

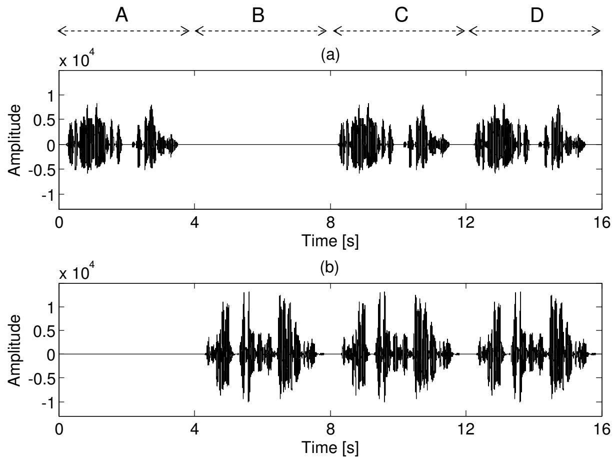

Fig. 2. (a) Received speech signal (female speech) and (b) near-end speech signal (male speech).

and the echo-power level was rapidly increased by about 20 dB. These simulated the residual echoes obtained from ADF process before and after the echo-path change, which means after and before the ADF convergence.

Microphone signal $y(n)$ is plotted in Fig. 3 (a). Figures 3 (b) and (c) plot the send signal after processing with conventional methods for time periods $L = 10$ and $L = 200$. The conventional method was calculated both from short and long time periods to compare the assessments. Figure 3 (d) shows the send signal with the proposed method. Time periods M and L with the proposed method were set at 10 and 200 according to the conventional method. As can be seen in Figs. 3 (a), (b), (c), and (d), the conventional and proposed methods sufficiently suppressed echo signals in the received single-talk situation (period A). Echo return loss enhancements (ELREs) were 27.27 dB and 27.13 dB with the conventional methods for $L = 10$ and $L = 200$, respectively. The ERLE of 27.33 dB was achieved with the proposed method.

However, these results indicate that the conventional method for $L = 10$ seemed to result in speech distortions in double-talk situations (periods C and D) compared with the conventional and proposed methods for $L = 200$. This is because the time period to estimate ACL was too short. In contrast, the conventional method for $L = 200$ was clearly evident from the insufficient echo suppression during the period from 12 s to 13 s immediately after the echo-path change because the time was prolonged, and as a result the tracking speed was slow. However, the proposed method was able to reduce the echo immediately after the change and retained the waveform of the near-end speech signal.

4.2. Subjective results

A multi-stimulus test with hidden reference and anchor (MUSHRA) using a 100-point scale, compliant with ITU-R BS.1534-1 [18], was used to test the quality of speech. All reference and evaluation signals are played to both ears with headphones (Sennheiser HD 280 Pro). Eight experienced listeners evaluated speech under three conditions: the send signal of conventional methods for $L = 10$ and $L = 200$, and the proposed method. The test results comparing the conventional and proposed methods are shown in Fig. 4. The vertical lines in the figure denote a 95% confidence interval. For each double-talk period (periods C and D), mean scores were awarded for four sound signals by eight listeners. As these results indicate, the proposed method had about the same quality as the conventional method for $L = 200$ during the period C, but a better score was observed in period D by using the new method that corrected the echo power level instantaneously. The new method for period D improved the sound quality by about eight points on a 100-point scale compared with the conventional method calculated for the same time period ($L = 200$), and a significant improvement was confirmed.

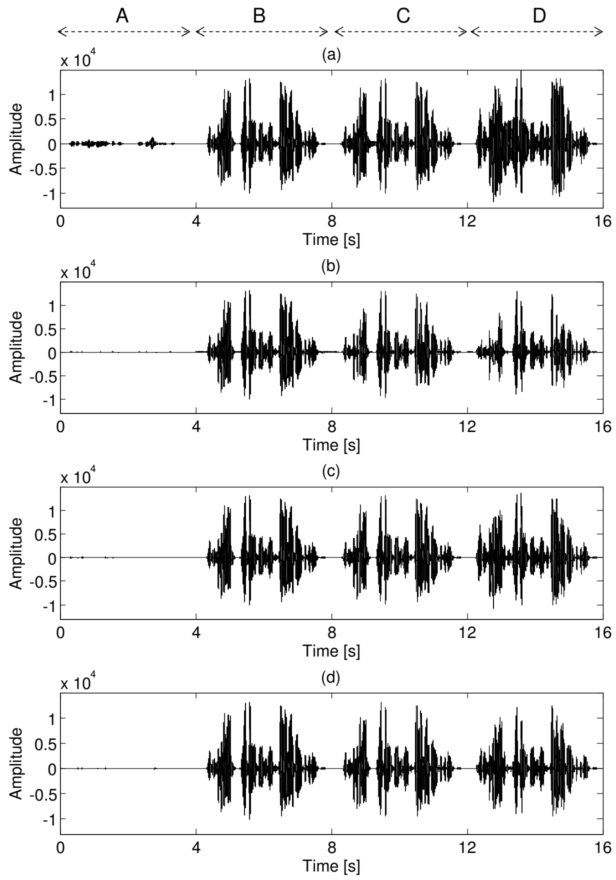

Fig. 3. (a) Microphone signal, (b) send signal with conventional method ($L = 10$), (c) send signal with conventional method ($L = 200$), and (d) send signal with proposed method.

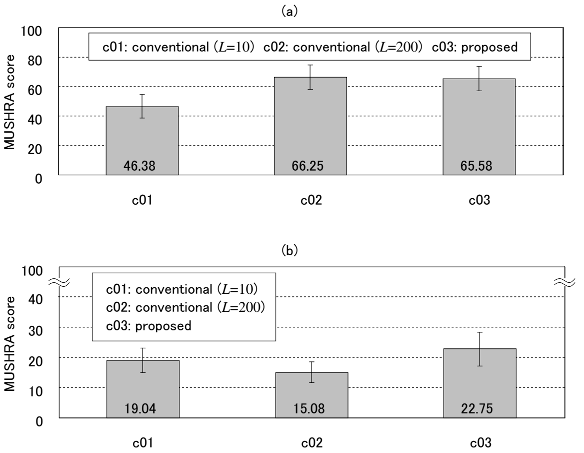

Fig. 4. Double-talk quality assessments for (a) period C and (b) period D.

5. CONCLUSION

A new method of instantaneously adjusting the echo power level for echo reduction was proposed. The echo power level was instantaneously estimated by focusing on all the frequency spectral components of the short-time power spectra of the error signal and the estimated residual echo from FDCC. Consequently, the residual echo was properly suppressed even immediately after a rapid change in the echo path. The experimental results revealed that the proposed method outperformed the conventional method during the abrupt increase of the echo-power level, and improved the quality of speech after the echo-path change by about eight points in the MUSHRA test.

6. REFERENCES

[1] J. Nagumo and A. Noda, “A learning method for system identification,” *IEEE Trans. Autom. Control*, vol. 12, no. 3, pp. 282–297, June 1967.

[2] S. Haykin, “Adaptive filter theory,” Third Edition, *Prentice-Hall, Inc.*, New Jersey, 1996.

[3] Suehiro Shimauchi, Yoichi Haneda, and Akitoshi Kataoka, “A robust NLMS algorithm for acoustic echo cancellation,” *IEICE Trans. Fundamentals*, vol. J89-A, no. 8, pp. 926–934, Aug. 2005.

[4] E. Hänsler and G. U. Schmidt, “Hands-free telephones — joint control of echo cancellation and postfiltering,” *Signal Process.*, vol. 80, no. 11, pp. 2295–2305, Nov. 2000.

[5] C. Avendano, “Acoustic echo suppression in the STFT domain,” *IEEE Workshop Sig. Process. to Audio and Acoust.*, vol. 21, no. 24, pp. 175–178, Oct. 2001.

[6] C. Faller and C. Tournery, “Robust acoustic echo control using a simple echo path model,” *Proc. ICASSP2006*, vol. 5, pp. 281–284, May 2006.

[7] Y. S. Park and J. H. Chang, “Frequency domain acoustic echo suppression based on soft decision,” *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 53–56, Jan. 2009.

[8] M. Fukui, S. Suehiro, A. Nakagawa, Y. Haneda, and A. Kataoka, “Acoustic-coupling level estimation for performance improvement of echo reduction,” *Proc. IWAENC2008*, Sep. 2008.

[9] M. Fukui, A. Nakagawa, S. Suehiro, Y. Haneda, and A. Kataoka, “Accurate echo power estimation for echo reduction,” *Proc. of the 2009 IEICE General Conference*, A-4-18, Mar. 2009.

[10] M. Fukui, A. Nakagawa, S. Suehiro, and Y. Haneda, “Echo reduction using Wiener gains considering short-time correlation between echo and near-end speech,” *Proc. IWAENC2012*, Sep. 2012.

[11] A. Favrot, C. Faller, and F. Kuech, “Modeling late reverberation in acoustic echo suppression,” *Proc. IWAENC2012*, Sep. 2012.

[12] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, “New optimal filtering approaches for hands-free telecommunication terminals,” *Signal Process.*, vol. 64, no. 1, pp. 33–47, Jan. 1998.

[13] S. Sakauchi, A. Nakagawa, Y. Haneda, and A. Kataoka, “Implementing and Evaluating of an Audio Teleconferencing Terminal with Noise and Echo Reduction,” *Proc. IWAENC2003*, pp. 191–194, Sep. 2003.

[14] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[15] J.S. Lim and A.V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Process. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[16] R. Le Bouquin Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, “Combined noise and echo reduction in hands-free systems: a survey,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[17] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” *Proc. ICASSP ’ 79*, vol.4, pp. 208–211, Apr. 1979.

[18] ITU-R Recommendation BS.1534-1, “Method for the subjective assessment of intermediate quality level of coding systems,” 2003.