A Comparison of Instrumental Measures for Wideband Speech Quality Assessment of Hands-free Systems in Echoic Condition

K. Steinert[1], S. Suhadi[2], and T. Fingscheidt[2]

[1] *Siemens AG, Corporate Technology, 81739 Munich, Germany, Email: kai.steinert.ext@siemens.com*

[2] *TU Braunschweig, Institute for Communications Technology, 38106 Braunschweig, Germany, Email: {s.suhadi,t.fingscheidt}@tu-bs.de*

Abstract

Instrumental speech component quality assessment presumes the availability of the processed speech portion of the enhanced output signal mixture along with its clean input counterpart. As the filtered speech component is unavailable in the context of black box measurements of speech enhancement systems, we have proposed a signal separation technique in earlier publications along with instrumental and subjective evaluations for the narrowband case. In this paper we apply the technique for a black box speech component quality assessment of two wideband hands-free systems with some instrumental measures. In particular, the echo reduction performance in the double-talk case will be investigated. Our findings are compared with the results of white box measures and of a subjective listening test.

Introduction

An important issue in instrumental quality assessment of hands-free systems, in the following also referred to as speech enhancement systems, is the evaluation of the speech component preservation. Usually, only the enhanced signal *mixture* is available, containing filtered speech, filtered/residual noise, and filtered/residual echo components. In the simulation environment the filtered speech may be obtained by the clean speech input signal being subject to the same processing as is applied to the input signal mixture (microphone signal). In cases of spectral weighting, the same gain multiplication can be applied to the clean speech input signal due to linearity. However, if the system under test also exhibits an echo canceler, the cancelation effect on the mere clean speech input signal, e.g., speech distortion in double-talk, cannot easily be calculated using the signals given, and is thus sometimes just ignored [1]. In any case, the quality of the speech portion can then be instrumentally evaluated in relation to its clean input counterpart.

We call these procedures *white box* tests, as opposed to a *black box* test, where the internal signal processing is entirely unknown and the method described consequently cannot be applied. Even if the microphone signal components (i.e., near-end speech signal, noise signal, and echo signal) are known, in general, the black box system output enhanced signal components (i.e., filtered speech, residual noise, and residual echo components) cannot easily be extracted from the enhanced signal. Therefore, we have proposed a technique, in the following termed signal separation, which allows to approximately compute the filtered speech component, given the microphone signal components and the enhanced signal mixture [2, 3]. This is even possible in the presence of nonlinearities as caused, e.g., by an echo canceler. Thus a black box test can be performed by assessing the filtered speech component quality with the near-end speech signal as reference. The applicability of the signal separation technique for 8 kHz sampled narrowband speech enhancement systems, which is also part of the new ITU-T Recommendation P.1100 [4, Sec. 8], has been demonstrated with several subjective and instrumental tests [5, 6].

In this contribution, we extend the application of the signal separation method to 16 kHz wideband systems. Parameters for that case, found by instrumental and informal subjective tests, are provided. Furthermore simulations were conducted to compare the performance of two wideband systems with respect to the speech component quality evaluated by black box instrumental measurements in relation to results of white box instrumental measurements. The performance was evaluated by employing the cepstral distance [7], the Itakura-Saito distance [8], and the MOS-LQO measure [9] for either evaluation setup. We carried out a subjective listening test and employed the results, all for the echoic condition in double-talk, as the reference for black box instrumental measurements.

White Box Test Setup

Figure 1 shows the speech enhancement processing setup for white box instrumental measurements as proposed in [1]. The system comprises an echo canceler with echo path model filter $\hat{h}(n)$ and a spectral weighting postfilter $g(n)$ (here it is written in time domain) for residual echo suppression (RES) and noise reduction (NR). The far-end excitation signal is denoted by $x(n)$ and the microphone signal by $y(n)$. The latter is composed of the far-end echo signal $d(n)$, the near-end speech signal $s(n)$, and the ambient noise signal $n(n)$:

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

The echo canceler output signal, the estimation error

$$e(n) = y(n) - \hat{d}(n), \tag{2}$$

is then computed with $\hat{d}(n) = x(n) * \hat{h}(n)$ being the echo signal estimate. After postfiltering the output signal $\hat{s}(n)$ is obtained.

If the internal processing of the system is unknown, i.e., in case of a black box system, only the enhanced signal $\hat{s}(n)$ can be observed together with the signal $y(n)$ and its corresponding components $\{d(n), s(n), n(n)\}$. Without the

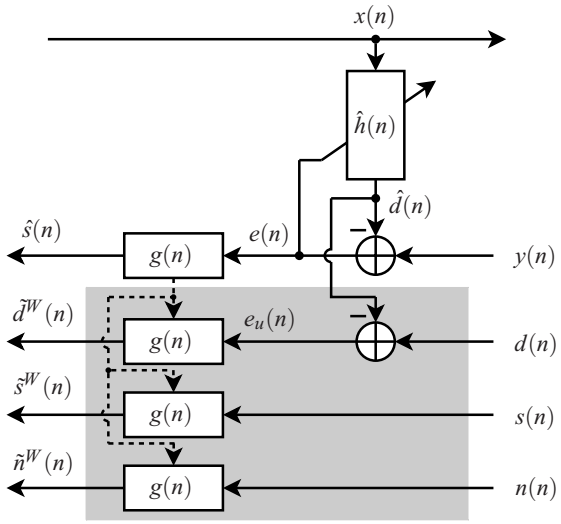

Figure 1: White box test setup of a system comprising echo cancelation and postfiltering (cf. [1]). The dashed lines denote a coefficient transfer. The enhanced signal components are generated in the gray box.

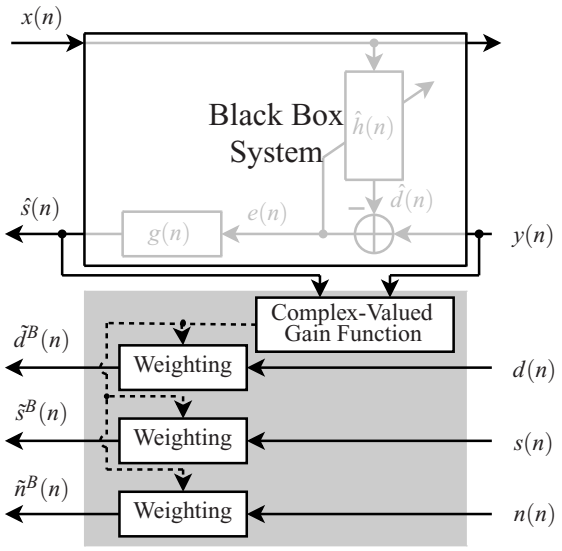

Figure 2: Black box test setup of an unknown system, modeled by complex-valued weighting with overlap-add according to signal separation technique (gray box, cf. [3]).

knowledge of the internal processing it would be unclear how, and to what extent, the speech, echo, and noise signals are modified by the system. On the contrary, in white box instrumental measurements, the components of the enhanced signal can be computed by performing the identical processing, which is computed using the microphone signal $y(n)$, to the microphone signal components separately. As for the postfilter $g(n)$, the filtering operation is performed on all microphone signal components by means of a coefficient transfer. However, the open question is how the three individual microphone signal components are modified by the subtraction operation (2) of the echo canceler. In [1] this subtraction is applied to the echo signal only, that is,

$$e_u(n) = d(n) - \hat{d}(n), \tag{3}$$

whereas the other microphone signal components are simply assumed not to be affected, as shown in Figure 1. Such assumption of course does not really model the reality—however, it is the only proposal known to the authors to achieve output signal components in speech enhancement systems comprising acoustic echo cancelation. This results in the enhanced signal components $\{\tilde{d}^W(n), \tilde{s}^W(n), \tilde{n}^W(n)\}$ of white box processing (here denoted by the superscript ^{W}) with

$$\hat{s}(n) = \tilde{d}^W(n) + \tilde{s}^W(n) + \tilde{n}^W(n). \tag{4}$$

The components in (4) could then be instrumentally evaluated with respect to those in (1).

Black Box Test by Signal Separation

With the signal separation technique the filtered version of the microphone signal components $d(n)$, $s(n)$, and $n(n)$ can approximately be calculated, given the input microphone signal components and the output signal mixture $\hat{s}(n)$ [2, 3]. Therefore it is assumed that the—potentially nonlinear and time-variant—speech enhancement processing can be modeled by short-term Fourier transform and spectral weighting with weights obtained from a complex-valued gain function (see gray box in Figure 2). For a detailed description the reader is referred to [2, 3, 6].

While our earlier papers on the topic dealt with narrowband system evaluation, we here consider the assessment of speech enhancement systems at a 16 kHz sampling rate. Therefore the signal separation parameters for the wideband case had to be found. The best results with respect to perceived similarity between the separated black box components and the respective components of the output signal mixture were obtained for an FFT length of 1024 samples, a frame shift of not more than 128 samples (i.e., twice the value of the 8 kHz parameters), and a Blackman window. A necessary condition for the signal separation to yield useful results is that

$$\hat{\hat{s}}(n) = \tilde{d}^B(n) + \tilde{s}^B(n) + \tilde{n}^B(n) \approx \hat{s}(n), \tag{5}$$

i.e., the sum of all enhanced signal components (according to Figure 2) approximates the output mixture signal, cf. (4). The superscript ^{B} of the enhanced signal components denotes the results of the signal separation technique applied to a black box system. Using the signal-to-approximation-error ratio (SAER), defined for a segment l with N samples as

$$SAER(l) = 10\ \log_{10}\left[\frac{\sum_{\nu=0}^{N-1} \hat{s}^2(\nu + lN)}{\sum_{\nu=0}^{N-1} \varepsilon^2(\nu + lN)}\right] \tag{6}$$

$$\varepsilon(n) = \hat{s}(n) - \hat{\hat{s}}(n), \tag{7}$$

we evaluated the appropriateness of approximation (5). The average SAER was calculated over all segments and files of an 8 kHz and a 16 kHz database of echoic and noisy speech signals. Each database consisted of 480 signals at various echo and noise levels. Zero, infinite, or undefined SAER values were discarded. Setting the SAER segment length

$N = 100$ for the narrowband and $N = 200$ for the wideband signals resulted in 31.3 dB SAER for the 8 kHz (note the similarity to an analog test in [5]) and 34.3 dB SAER for the 16 kHz sampling rate.

Experimental Setup

The application of the signal separation technique to wideband systems is demonstrated by the performance evaluation of two 16 kHz speech enhancement systems w.r.t. the filtered speech component quality. For this purpose, the filtered speech component in the time domain, $\{\tilde{s}^B(n), \tilde{s}^W(n)\}$, serves for our instrumental and subjective quality measurements. System A comprises a frequency domain adaptive filter with model-based step size control (realization of [10] at 16 kHz sampling rate) for echo cancelation and an NR/RES Wiener postfilter. System B consists of a filterbank acoustic echo canceler with near-optimum step size control and an a priori SNR-driven Wiener filter for NR/RES (16 kHz version of [11]). The systems are evaluated subjectively and instrumentally after the initial convergence has taken place.

The input signals to the speech enhancement systems were generated synthetically using an American English speech database and 10 different car impulse responses at a 16 kHz sampling rate. Speech of four male and four female speakers was used for the far-end and near-end speech signals. After filtering each far-end signal with one of the impulse responses mentioned, the near-end speech was added to obtain signal-to-echo ratios (SERs) of 0, 5, and 10 dB. The SER was determined as the ratio of the power of $s(n)$ to that of $d(n)$. Therefore we employed the active speech level for the power estimation, using the ITU-T software tool library [12]. Altogether we obtained 120 different (loudspeaker and microphone) input signal pairs.

Instrumental Assessment

For the instrumental evaluation of the speech component of the hands-free systems we have considered the wideband MOS-LQO according to [9] and two LPC-based measures. As for the latter two, we assume that the speech signal $s(n)$ can be approximated in the short term by an allpole filter of pth order according to

$$s(n) = \sum_{i=1}^{p} a_s(i)s(n-i) + G_s u(n) \tag{8}$$

with the LPC coefficients $\mathbf{a}_s = [1, -a_s(1), -a_s(2), \ldots, -a_s(p)]^{\mathrm{T}}$, the filter gain $G_s = \sqrt{\mathbf{r}_s^{\mathrm{T}}\mathbf{a}_s}$, the unit variance white noise signal $u(n)$, and the autocorrelation vector $\mathbf{r}_s = [r_s(0), r_s(1), \ldots, r_s(p)]^{\mathrm{T}}$ with $r_s(i) = \mathrm{E}\{s(n)s(n-i)\}$. We define a symmetric Toeplitz autocorrelation matrix $\mathbf{R}_s$ whose first column is given by $\mathbf{r}_s$. These quantities are all calculated segment by segment for an entire signal and $\mathbf{a}_s$ is obtained from the Levinson-Durbin algorithm [13]. The parameters for the filtered speech component $\tilde{s}(n) = \{\tilde{s}^B(n), \tilde{s}^W(n)\}$ are obtained in the same way. The Itakura-Saito distance is then given as [8]

$$d_{IS}(\mathbf{a}_s, \mathbf{a}_{\tilde{s}}) = \frac{G_s}{G_{\tilde{s}}} \frac{\mathbf{a}_{\tilde{s}}^{\mathrm{T}}\mathbf{R}_s\mathbf{a}_{\tilde{s}}}{\mathbf{a}_s^{\mathrm{T}}\mathbf{R}_s\mathbf{a}_s} + \log(\frac{G_{\tilde{s}}}{G_s}) - 1 \tag{9}$$

and the cepstral distance is calculated as [7]

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k}, \quad 1 \le m \le p \tag{10}$$

$$d_{CD}(\mathbf{c}_s, \mathbf{c}_{\tilde{s}}) = \frac{10}{\ln 10} \sqrt{2 \sum_{k=1}^{p} [c_s(k) - c_{\tilde{s}}(k)]^2}. \tag{11}$$

These segmental results are averaged over the whole signal. In all cases, the speech component of the output signal, $\tilde{s}(n)$, was assessed relative to that of the input signal, $s(n)$, using the respective black box and white box instrumental measurement setup. The average results of system A were subtracted from those of system B to obtain difference values of the instrumental measures. These were averaged over all signals of the database for each SER, yielding measures in the following referred to as ΔMOS-LQO, ΔCD, and ΔIS.

Subjective Assessment

In a subjective listening test, 16 test persons (experts and non-experts) had to rate the quality of the speech component in system B output with respect to that in system A output. The test results are reported in terms of the comparison mean opinion scores (CMOS) [14] ranging in 7 steps from -3 (much worse signal component quality in system B) over 0 (about the same as system A) to $+3$ (much better signal component quality in system B). The subjective test is based on a subset of the database used for the instrumental test.

Experimental Results

The instrumental and subjective measurement results with the respective two-sided 95% confidence intervals are given in Table 1. The reader is reminded to the fact that all values are actually differences of system B performance relative to system A performance. That is, a positive distance (ΔCD and ΔIS) and a negative ΔMOS-LQO and CMOS would all suggest a preference of system A w.r.t. the speech component quality.

In fact, for all input SERs, the ΔMOS-LQO and CMOS values have a negative sign and the distance values have a positive (i.e., the opposite) sign. The subjective CMOS results for the three SER conditions, e.g., are between -1.13 and -1.26, meaning that the speech component quality of system B was auditively perceived a little more than "slightly worse," though not yet "worse" (which would have been rating -2), compared to system A. This was indeed measured consistently over all instrumental quality measures for black box and white box tests and the subjective test. However, two ΔIS values exhibit a relatively large confidence interval, and thus are not clearly positive in the sense of statistical significance. Nevertheless, the tendency of the black box instrumental measurement toward a preference of system A with the signal separation technique applied to 16 kHz systems is consistently confirmed by all white box instrumental measurements and, most importantly, the subjective test.

Measure	SER [dB]		
	0	5	10
Black box ΔCD	$+0.47 \pm 0.07$	$+0.53 \pm 0.07$	$+0.71 \pm 0.07$
White box ΔCD	$+1.02 \pm 0.08$	$+1.05 \pm 0.08$	$+1.24 \pm 0.08$
Black box ΔIS	$+2.61 \pm 0.72$	$+0.64 \pm 0.73$	$+0.01 \pm 0.79$
White box ΔIS	$+2.18 \pm 0.78$	$+0.96 \pm 0.78$	$+0.98 \pm 0.83$
Black box ΔMOS-LQO	-0.23 ± 0.03	-0.29 ± 0.02	-0.34 ± 0.02
White box ΔMOS-LQO	-0.29 ± 0.06	-0.41 ± 0.06	-0.53 ± 0.07
CMOS (subj. test)	-1.13 ± 0.19	-1.26 ± 0.20	-1.26 ± 0.19

Table 1: Speech component quality evaluation results for echoic condition: performance of system B relative to system A

Furthermore, comparing the SER conditions from 0 to 10 dB, there is a consistent tendency on the one hand in white box, black box, and CMOS results, on the other hand also in ΔMOS-LQO, ΔCD, and CMOS measures. Only the ΔIS measure turns out not to be applicable to our task—it shows a quite different tendency as, e.g., CMOS. In general we can conclude that a black box ΔMOS-LQO test nicely reflects the CMOS measure.

Such comparative quality assessment results of two speech enhancement systems are of major practical interest, especially if only black box instrumental measurement tests are feasible. In our paper we applied the earlier-proposed signal separation technique to the speech component quality assessment of wideband systems, demonstrating the same comparison results of instrumental measures such as MOS-LQO or cepstral distance in a black box test scenario as in the subjective CMOS test.

Conclusion

We have considered the problem of speech enhancement system evaluation for the wideband case. An instrumental black box test method, originally proposed for the narrowband case, was assessed to compare two 16 kHz systems w.r.t. the speech component quality of echoic signals. We have shown the similarity of results from the wideband application of the convenient black box instrumental measurement test method with white box instrumental measurement of MOS-LQO or cepstral distance and subjective tests. We conclude that the convenient black box MOS-LQO instrumental measure best reflects the subjective CMOS results—both in answering the question: which system is better, and also in giving information over different signal-to-echo ratios (SER).

Acknowledgment

The authors are grateful to Carsten Last for carrying out extensive simulations.

References

[1] S. Gustafsson, R. Martin, and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-free Telephony," Elsevier Signal Processing, vol. 64, 1998, pp. 21-32.

[2] T. Fingscheidt and S. Suhadi, "Experiments on Speech, Noise, and Echo Separation for Quality Assessment of Hands-free Systems," in Proc. of DAGA07, Stuttgart, Germany, Mar. 2007.

[3] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in Proc. of INTERSPEECH07, Antwerp, Belgium, Aug. 2007.

[4] "ITU-T Rec. P.1100, Narrowband Hands-free Communication in Motor Vehicles," ITU, Oct. 2008.

[5] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach," in Proc. of ICASSP08, Las Vegas, NV, USA, Apr. 2008.

[6] K. Steinert, S. Suhadi, T. Fingscheidt, and M. Schönle, "Instrumental Speech Distortion Assessment of Black Box Speech Enhancement Systems," in Proc. of IWAENC 08, Seattle, WA, USA, Sep. 2008.

[7] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," IEEE Journal on Selected Areas in Communications, vol. 6, no. 2, Feb. 1988, pp. 242-248.

[8] B.-H. Juang, "On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation," AT&T Bell Laboratories Technical Journal, vol. 63, no. 8, Oct. 1984, pp. 1477-1498.

[9] "ITU-T Rec. P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.

[10] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-free Telephones," Elsevier Signal Processing, vol. 86, no. 6, 2006, pp. 1140-1156.

[11] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction," in Proc. of ICASSP08, Las Vegas, NV, USA, Apr. 2008.

[12] "ITU-T Rec. G.191, Software Tools for Speech and Audio Coding Standardization," ITU, Sept. 2005.

[13] J. Durbin, "The Fitting of Time-Series Models," Revue Inst. Int. de Stat., v. 28, no. 3, 1960, pp. 233-243.

[14] "ITU-T Rec. P.800, Methods for Subjective Determination of Transmission Quality," ITU, Aug. 1996.