INSTRUMENTAL SPEECH DISTORTION ASSESSMENT OF BLACK BOX SPEECH ENHANCEMENT SYSTEMS

Kai Steinert[1]*, Suhadi Suhadi*[2]*, Tim Fingscheidt*[2]*, and Martin Schönle*[1]

[1] Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81739 Munich, Germany
Email: {kai.steinert.ext,martin.schoenle}@siemens.com
[2] TU Braunschweig, Institute for Communications Technology,
Schleinitzstr. 22, 38106 Braunschweig, Germany, Email: {s.suhadi,t.fingscheidt}@tu-bs.de

ABSTRACT

An important parameter in quality assessment of speech enhancement systems is speech distortion, measured in terms of quality of the speech component. In fact, in the context of noise reduction, the user tends to prefer a certain degree of residual noise over distorted speech with suppressed background noise. The challenge of instrumental speech component quality evaluation lies, among others, in the mere availability of the enhanced output signal mixture rather than its speech portion. In this paper we present a method to extract the speech component from the enhanced output signal with high accuracy, given the input signal components speech, noise, and echo. We apply this method to a black box speech component quality comparison of two speech enhancement systems and report on instrumental and subjective tests with focus on double-talk.

Index Terms— Instrumental speech quality assessment, non-blind signal decomposition, speech enhancement

1. INTRODUCTION

In algorithmic development of noise reduction schemes, researchers have a convenient method of instrumental speech component quality assessment at their disposal. The noisy input signal can be constructed as the sum of a clean speech and a noise signal. Assuming spectral weighting in the short-time Fourier transform (STFT) domain as is typical for noise reduction systems, the weights which are applied to the noisy input signal can be logged and applied separately to the clean speech input signal component to obtain the processed speech component [1]. Clearly, as this is a highly intrusive approach, it is feasible only if the internal processing of the speech enhancement system is known with its parameters such as frame length, frame shift, window function, and the weights. We call this approach a *white box* test. While this approach is perfectly applicable for frequency domain filtering such as noise reduction and residual echo suppression, it has drawbacks concerning time domain acoustic echo cancellation: In [1] the subtraction of the estimated echo is modeled as affecting solely the echo signal component of the microphone signal. Unrealistically, the speech component is then only affected by the residual echo suppression and noise reduction, which are implemented as a postfilter after the echo canceller.

Here we are interested either in such sophisticated speech enhancement systems whose effect on the speech signal cannot simply be described by logged spectral weights, or in unknown (e.g., hardware) systems—commonly referred to as *black box* systems. In these cases, we merely have the output signal mixture to measure the speech component quality.

We previously suggested a technique which enables us to extract the processed speech component from the enhanced output signal with high accuracy, given the clean speech component [2–4]. This way we are able to judge the impact of the speech enhancement system on the speech portion only. The methodology assumes the black box speech enhancement system to exhibit a digital I/O, and has been included by ITU-T SG12 into the new draft P.1100 recommendation on hands-free communication in motor vehicles [5]. In the paper at hand, we compare instrumental white box and black box measurement results with subjective listening test findings for the double-talk case. We use the MOS-LQO measure [6], which can be applied with good accuracy due to the availability of the speech portion of the output signal after having extracted the processed speech component.

2. DECOMPOSITION OF THE ENHANCED SIGNAL MIXTURE

In a black box test scenario of sophisticated speech enhancement systems comprising, e.g., noise reduction and echo cancellation, the internal processing usually is unknown. In order to allow for a later decomposition of the enhanced signal, we first acquire the three input signal components *clean speech* $s(n)$, *background noise* $n(n)$, and *echo* $d(n)$ as follows. We have to digitally record our near-end speech test signal $s(n)$ and the test noise $n(n)$ separately via the microphone and the A/D converter of the speech enhancement device under test

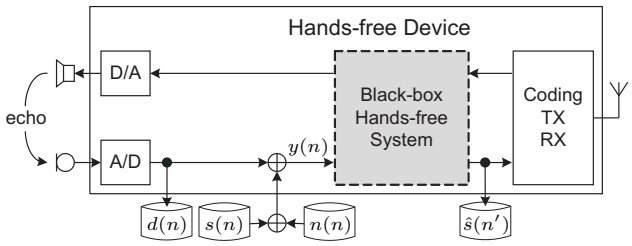

Fig. 1. Black box test of an arbitrary speech enhancement system including noise reduction and echo cancellation.

(see Fig. 1), and then add both signals. In the actual test of the speech enhancement system a far-end signal is to be fed into the downlink input of the system. In the loudspeaker-enclosure-microphone (LEM) system only the echo signal is captured by the microphone and digitally stored after A/D conversion. The pre-recorded near-end speech plus the noise of the respective test case are then to be added to the captured echo signal $d(n)$ and are input to the black box speech enhancement system in the uplink. Note that this separate acquisition of input signals models the microphone and the A/D converter as a linear system. In practice—given a proper scaling of signals—this turns out to be approximately a quite reasonable assumption.

Following this recording methodology, we can indeed observe the corresponding enhanced speech signal $\hat{s}(n')$—with sample index n' having a certain delay with respect to sample index n—and its input component signals: near-end speech signal $s(n)$, noise signal $n(n)$, and echo signal $d(n)$. The rest of our investigations is purely performed as offline processing based on the stored digital signals $d(n), n(n), s(n), \hat{s}(n')$. We are aware that such signals are often not yet digitally accessible in today's speech enhancement devices. However, they could be made accessible via a digital interface as it has been described in the new draft ITU-T Recommendation P.1100 [5]. For any speech enhancement software simulation these signals should be easily available.

Having available $s(n)$, $n(n)$, $d(n)$, and $\hat{s}(n)$ (after time delay compensation of $\hat{s}(n')$), we process each signal in a frame-wise manner with a blackman window, DFT length 512, and frame shift 64. Previous work [3] has shown that these settings yield the best performance for 8 kHz sampled signals. In the DFT domain

$$Y_l(k) = S_l(k) + N_l(k) + D_l(k) \tag{1}$$

holds and IDFT with overlap-add results in $y(n)$ again. Capital letters denote the DFT of the respective signals with the frame index l and the frequency bin k. Without loss of generality for the analysis to follow, assume now that $D_l(k)$ is already included in $N_l(k)$, so that we have a speech component and a noise component (which includes the echo component). The amplitude and phase formulations in the frequency domain are then

$$|Y_l(k)|e^{j\phi_{Y_l}(k)} = |S_l(k)|e^{j\phi_{S_l}(k)} + |N_l(k)|e^{j\phi_{N_l}(k)}. \tag{2}$$

We now simply model our possibly unknown, time-variant, and nonlinear speech enhancement system processing by assuming that it applies a *complex*-valued gain function $G_l(k) \in \mathbb{C}$ in our overlap-add framework according to

$$|\hat{S}_l(k)|e^{j\phi_{\hat{S}_l}(k)} = G_l(k) \cdot |Y_l(k)|e^{j\phi_{Y_l}(k)}. \tag{3}$$

Given (3), the complex gain of the speech enhancement system shall be computed by division according to

$$G_l(k) \approx \min\left[\frac{|\hat{S}_l(k)|}{|Y_l(k)|}, 1\right] \cdot \frac{e^{j\phi_{\hat{S}_l}(k)}}{e^{j\phi_{Y_l}(k)}}. \tag{4}$$

The $\min[\cdot]$ operation in (4) avoids audible artifacts that sound similar to musical noise. The filtered speech and noise components of the enhanced speech signal can be computed individually in the frequency domain by

$$|\tilde{S}_l(k)|e^{j\phi_{\tilde{S}_l}(k)} = G_l(k) \cdot |S_l(k)|e^{j\phi_{S_l}(k)} \quad \text{and} \tag{5}$$

$$|\tilde{N}_l(k)|e^{j\phi_{\tilde{N}_l}(k)} = G_l(k) \cdot |N_l(k)|e^{j\phi_{N_l}(k)}. \tag{6}$$

The limitation in (4) results in the sum of the filtered speech component and the filtered noise component in the frequency domain only approximating the enhanced speech signal as

$$|\tilde{S}_l(k)|e^{j\phi_{\tilde{S}_l}(k)} + |\tilde{N}_l(k)|e^{j\phi_{\tilde{N}_l}(k)} \approx |\hat{S}_l(k)|e^{j\phi_{\hat{S}_l}(k)}. \tag{7}$$

The approximation error, however, has been shown to be about -30 dB [4]. Eqs. (5) and (6) hold for an additive mixture of filtered speech and noise/echo components, as assumed before. The processed speech component in the time domain $\tilde{s}(n)$, which will be the subject of our following investigations, is computed by subsequent IDFT of (5) and overlap-add. It serves for our instrumental and subjective quality measurements.

3. EXPERIMENTAL SETUP

The application of the signal decomposition technique is demonstrated by the performance evaluation w.r.t. the speech component quality of two speech enhancement systems. System A comprises a time-domain NLMS echo canceller with VAD-controlled fixed-step sizes and a frequency domain noise reduction based on the least square amplitude estimator with VAD-based noise power estimation [7]. Meanwhile, system B consists of a filterbank acoustic echo canceller with near-optimum step size control and an a priori SNR-driven Wiener filter is applied as residual echo and noise reduction [8]. The systems are evaluated subjectively and instrumentally after the initial convergence has taken place.

The input data to the speech enhancement systems is generated synthetically using the NTT-AT speech and car noise

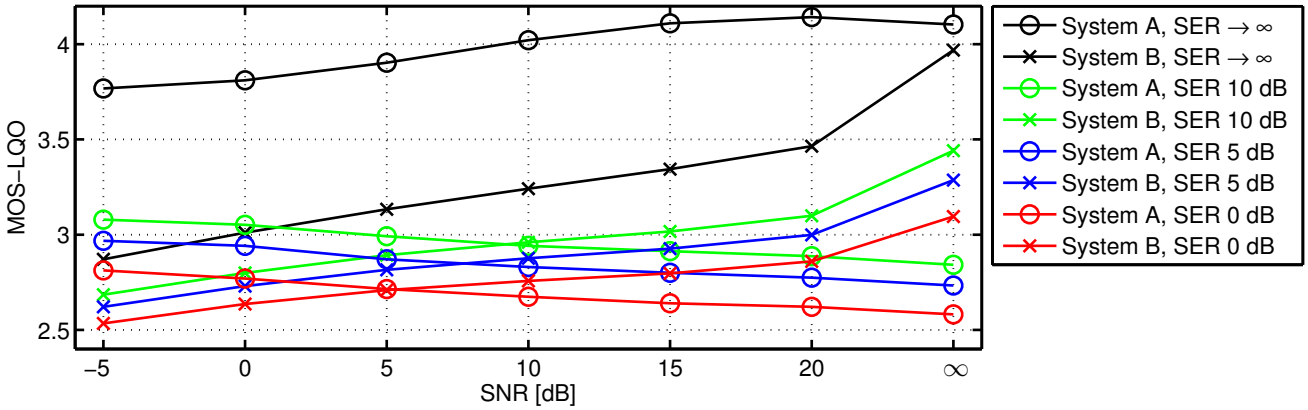

Fig. 2. Black box instrumental measurement results of speech component quality in terms of MOS-LQO: System A (circles) compared to system B (crosses) for various SNR and SER (different colors) conditions.

databases and a car impulse response at an 8 kHz sampling rate. Four male and four female speakers of American English were used for the far-end and near-end speech signals. However, of all possible combinations only 8 were chosen to obtain 2 far-end/near-end combinations of male-female, female-male, female-female, and male-male speakers. The far-end signal was filtered with the impulse response prior to the addition with the near-end speech. The signal-to-echo ratios (SERs) were 0, 5, 10, and ∞ dB, and 40 different noise files were added with the signal-to-noise ratio conditions (SNRs) of -5, 0, 5, 10, 15, 20, and ∞ dB. Altogether we obtained 1120 different disturbed input (loudspeaker and microphone) signal pairs.

The SNR was determined as the ratio of the power of $s(n)$ to that of $n(n)$. The SER was calculated as the ratio of the power of $s(n)$ and that of $d(n)$. In all cases we employed the active speech level for the power estimation, using the ITU-T software tool library [9].

3.1. Instrumental Assessment

We are interested in the quality of the enhanced output speech component $\tilde{s}(n)$. The subjective quality of $\tilde{s}(n)$ relative to the clean speech input $s(n)$ is predicted instrumentally using PESQ MOS scores [10], mapped to the MOS-LQO scale [6]. The PESQ measure compensates for a possible time lag and a broadband amplitude scaling between both signals under consideration. It is averaged over all signals of the database for each SNR and SER condition.

To facilitate contrasting the instrumental to the subjective results described below, we calculated the difference of each measured value between system B and system A, in the following referred to as ΔMOS-LQO.

3.2. Subjective Assessment

In a subjective listening test we assessed the echo-free (SER $\rightarrow \infty$) and the noise-free (SNR $\rightarrow \infty$) case separately. In both cases, 16 listeners (experts and non-experts) had to rate the quality of the speech component in system B with respect to that in system A. The test results are reported in terms of the comparison mean opinion scores (CMOS) [11] 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).

4. EXPERIMENTAL RESULTS

The black box instrumental measurement results are depicted in Fig. 2. The abscissa indicates the SNR, while the different colors of the curves stand for the various SER conditions, and the markers indicate the system. The ordinate represents the MOS-LQO scale.

It can be seen that, for finite SERs, system B yields higher MOS-LQO values (i.e., less speech distortion) for SNR $\geq$ 10 dB. For the echo-free case (SER $\rightarrow \infty$) and low SNRs system A exhibits better speech preservation.

For a subset of SNR and SER conditions, the black box ΔMOS-LQO values are shown in Tabs. 1 and 2, along with white box measurements. Concerning the sign, the ΔMOS-LQO values as computed in the black box test yield identical results to the white box test, clearly stating which system is better in which condition: System A turns out to have a noise reduction better preserving near-end speech (and clean speech performance with no echo), while according to Tab. 2 system B consistently proves to provide less speech distortions during acoustic echo cancellation.

The subjective test results can be seen in Figs. 3 and 4. The bar chart shows the CMOS scores of system B relative to system A, that is, the extent to which the speech component quality of system B is perceived as better (positive values) or worse (negative values) compared to system A, and the 95% confidence intervals. As is shown in Fig. 3, the quality of the speech component of system A is preferred to some extent for the echo-free case. Comparing the instrumental black box results with the subjective findings, we find that the negative

SNR [dB]	0	5	10
Black box ΔMOS-LQO	−0.80	−0.77	−0.78
White box ΔMOS-LQO	−0.31	−0.40	−0.54

Table 1. *Echo-free* case: Instrumental results of system B vs. system A, black box values taken from Fig. 2

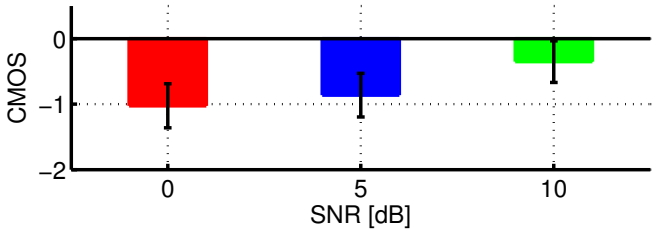

Fig. 3. *Echo-free* case: Subjective test results (CMOS) with 95% confidence intervals: speech component quality for system B vs. system A

SER [dB]	0	5	10
Black box ΔMOS-LQO	0.51	0.55	0.60
White box ΔMOS-LQO	1.40	1.40	1.38

Table 2. *Noise-free* case: Instrumental results of system B vs. system A, black box values taken from Fig. 2

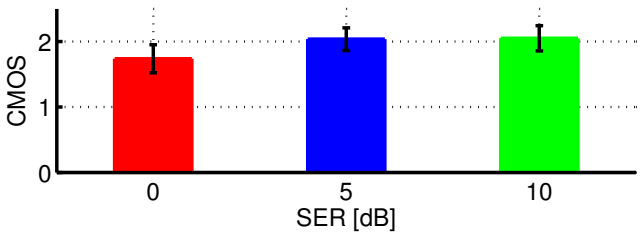

Fig. 4. *Noise-free* case: Subjective test results (CMOS) with 95% confidence intervals: speech component quality for system B vs. system A

ΔMOS-LQO values in Tab. 1 correspond to negative CMOS values in Fig. 3. Obviously, for our test, the black box signal decomposition has yielded similar relative (instrumentally measured) results as the white box test and as the subjective test.

The noise-free condition in Fig. 4 suggests a preference of system B over A. The positive ΔMOS-LQO values in Tab. 2 clearly correspond to the positive CMOS values. It should be noted that the confidence intervals in the CMOS scores do not allow the statement on SNR dependency (Fig. 3), or SER dependency (Fig. 4). A statistically reliable conclusion, however, is that system A has a noise reduction better preserving the quality of the speech component, while system B offers an acoustic echo compensation with less speech distortion. Such comparative quality assessment results of two speech enhancement systems are of major practical interest, especially if only black box tests are possible. In our paper we presented an instrumental path towards achieving the same comparison results as in the subjective CMOS test.

5. CONCLUSION

In this paper we have presented an instrumental speech component quality assessment method for black box speech enhancement systems. In the most difficult test condition *double-talk* we have shown the similarity of results from (a) our new convenient black box instrumental test method with (b) instrumental white box tests, and (c) subjective tests. Instrumental measurement of the speech component by MOS-LQO during double-talk using our technique is an optional test case in ITU-T's new draft recommendation P.1100 [5]. At the workshop, speech samples obtained by the signal decomposition technique from section 2 will be presented.

6. REFERENCES

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

[2] T. Fingscheidt and S. Suhadi, "Experiments on Speech, Noise, and Echo Separation for Quality Assessment of Hands-free Systems," in *Proc. of DAGA'07*, 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 INTERSPEECH'07*, Antwerpen, Belgium, Aug. 2007.

[4] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach," in *ICASSP'08*, Las Vegas, Nevada, USA, Apr. 2008.

[5] "ITU-T Draft Recommendation P.1100, Narrowband Hands-Free Communication in Motor Vehicles," ITU, June 2008.

[6] "ITU-T Recommendation P.862.1, Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO," ITU, Nov. 2003.

[7] M. Schönle, C. Beaugeant, K. Steinert, H.W. Löllmann, B. Sauert, and P. Vary, "Hands-Free Audio and Its Application to Telecommunication Terminals," in *Proc. of AES 2006*, Seoul, Korea, Sept. 2006.

[8] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction," in *ICASSP'08*, Las Vegas, Nevada, USA, Apr. 2008.

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

[10] "ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ)," ITU, Feb. 2001.

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