A New Methodology for Objective Performance Assessment of Hands-free Systems in Double-Talk

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

[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

Speech quality evaluation of hands-free terminals is a complex task. Several aspects have to be taken into account such as the various conversational situations and a possibly nonlinear and time-variant system behavior. The lack of access to the internal signal processing of black box systems complicates a separate assessment of the processed clean speech, echo, and noise. In this paper we present an objective evaluation of the performance of two hands-free systems in terms of echo attenuation and speech distortion during double-talk. Based on an earlier published signal separation method, we consider the processed echo and the processed clean speech relative to the respective unprocessed signal individually. Our findings are compared with the results of a subjective listening test.

Introduction

Quality measurements of speech enhancement systems can be time-consuming and costly. In algorithmic research and development, where the system under consideration is regarded as a white box (i.e., internal signals and states are accessible), the effect of the processing on the microphone path input signal components (i.e., far-end speaker's echo, near-end speech, and noise) can in principle be examined by separately processing the respective input signal components with the same framing and spectral weighting. This would then allow for a separate evaluation of each signal component. However, this method is not applicable for black box systems, in which case the internal processing is unknown. In this contribution we examine an earlier published signal separation method [1, 2] that allows to approximately compute the processed speech, noise, and echo parts of the microphone path output signal for the purpose of their individual evaluation. The methodology assumes digital access to the black box system (as is under consideration in ITU-T's SG12 focus group FITCAR (From/In/To Cars Communication)), and that the input signal was generated artificially by adding its components.

Specifically, the prerecorded local speech and noise signals and the echo signal picked up by the microphone are added in real-time to constitute the microphone input signal. At the same time, the enhanced system output signal in sending direction is recorded in the digital domain. The separate processed signal components are estimated offline subject to an approximate transfer function computed in the STFT domain describing the effect of the black box system. For details the reader is referred to [1, 2]. The data gathered that way then constitutes the basis for objective (and perhaps also subjective) assessment of speech distortion, noise and echo attenuation.

Experimental Setup

In this paper we apply the signal separation method to evaluate two hands-free systems ([3] and one similar to [4]—gain loss control not used), in the following referred to as system A and B, respectively. Our measures of interest, i.e., the quality degradation of the speech component (sending direction) and the echo attenuation of both systems, are first assessed objectively using the individual processed signal components, and subsequently compared with the results of a subjective listening test of the composite output signal. The systems are evaluated after the initial convergence has taken place.

The input data to the hands-free systems is generated synthetically using the NTT-AT speech and car noise databases and a car impulse response. 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, and 10 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. Thus we obtained 840 different disturbed input signals.

Simulation Results

The performance of the speech enhancement systems in terms of the PESQ MOS [5] and the SER improvement is measured objectively. Rather than the SER improvement, the (undisturbed) echo return loss enhancement (ERLE) could have been considered as well, as is often the case in the context of echo cancellation. However, it has not been applied here since it does not compensate for a possible gain variation taking place in the hands-free systems, and mistakenly being measured as an echo reduction. We estimate the SER improvement as the difference between the output and the input signal SER. These two measures are both averaged over all signals for

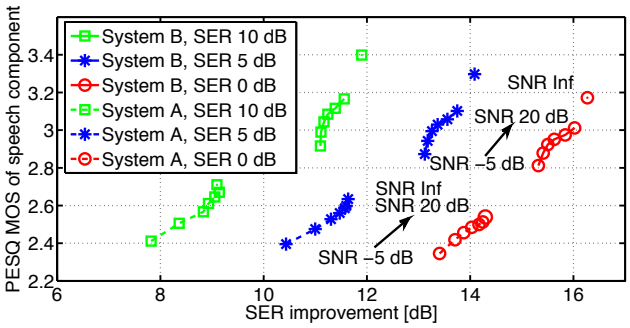

Figure 1: Objective measurement results

each SNR and SER condition.

The objective measurement results are depicted in Fig. 1. It can clearly be seen that system B has PESQ MOS values about 0.5 or more points higher than those of system A. For the SER improvement the results of system B are about 2-3 dB higher than those of system A. Furthermore it can be noticed that the noise-free case ("SNR Inf" in Fig. 1) yields a much higher PESQ MOS value for system B than the noisy cases, whereas the improvement for the noise-free case for system A is rather small.

The objective measures of the separate signal components give an important hint about the expected system performance and may prove useful for localizing the cause of quality impairments. Yet, it is to be proven that they correspond to the results of subjective quality assessments.

For that reason, we have conducted a subjective listening test using a tool similar to the one described in [6]. A subset of only noise-free samples was taken from our test database. 16 expert listeners had to rate if, and to what extent, they preferred either of the two systems over the other. Similar to the objective test, the quality of the speech component and the residual echo level (not the attenuation though) of system A and B have been compared in a CCR-like test (see Table 1).

CMOS	Quality of the speech component	Echo level
3	much better	much weaker
2	better	weaker
1	slightly better	slightly weaker
0	about the same	about the same
-1	slightly worse	slightly stronger
-2	worse	stronger
-3	much worse	much stronger

Table 1: The used CMOS rating scale

The subjective test results can be seen from the bar chart in Fig. 2. The chart depicts the CMOS values representing the preference of system B over system A colored according to the SERs as in Fig. 1, along with the respective 95% confidence intervals. The filled and the empty bars stand for the speech component quality and the echo level, respectively. The chart suggests a *significant* preference of system B over system A, both in terms of the speech quality and the echo level. The plot indicates that the signal quality of system B was judged

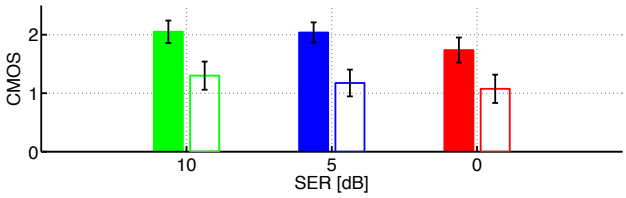

Figure 2: Subjective test results (CMOS) with 95% confidence intervals for the noise-free case: speech component quality (filled bars) and residual echo level (empty bars) for system B relative to system A

to be approximately "better" (rating 2) with the echo level being about "slightly weaker" (rating 1). This result corresponds to the objective measurement where system B turned out to exhibit higher PESQ MOS values and higher SER improvements. In particular, the decreasing CMOS value for the residual echo level towards lower SERs seems to confirm the objective results in Fig. 1 where (for SNR $\rightarrow \infty$) at SER = 10 dB system B is 3 dB better in terms of SER improvement, while at SER = 0 dB the difference is merely 2 dB. However, statistical significance has to be proven in further tests.

Conclusions

In this paper we have applied the earlier published signal separation method to objectively evaluate the quality of two hands-free systems. We measured the quality of the speech component and the echo attenuation during double-talk using the processed signal components of the enhanced output signal mixture. A CCR subjective listening test was conducted for the noise-free case and suggested a strong correspondence to the objective test results. Listening tests for other noise cases are subject of further studies. We conclude that the signal separation method appears to be a useful approach to obtain objective results of speech enhancement systems with unknown internal processing.

References

[1] Fingscheidt, T., Suhadi, S.: Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo. Proc. INTERSPEECH, Antwerp, Belgium, Aug. 2007

[2] Fingscheidt, T., Suhadi, S., Steinert, K.: Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach. Proc. ICASSP, Las Vegas, NV, USA, Mar. 2008

[3] Schönle, M. et al.: Hands-Free Audio and its Application to Telecommunication Terminals. Proc. AES, 29th conference, Seoul, South Korea, Sep. 2006

[4] Steinert, K., Schönle, M., Beaugeant, C., Fingscheidt, T.: Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction. Proc. ICASSP, Las Vegas, NV, USA, Mar. 2008

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

[6] Drascher, T., Gilg V., Schönle, M.: A New Tool for Evaluation of Sound Quality in Carkit Telephony. Proc. DAGA, Braunschweig, Germany, Mar. 2006