TOWARDS OBJECTIVE QUALITY ASSESSMENT OF SPEECH ENHANCEMENT SYSTEMS IN A BLACK BOX APPROACH

Tim Fingscheidt, Suhadi Suhadi

Institute for Communications Technology
Braunschweig Technical University
D – 38106 Braunschweig, Germany
{t.fingscheidt, s.suhadi}@tu-bs.de

Kai Steinert

Siemens AG
Corporate Technology
D – 81730 Munich, Germany
kai.steinert.ext@siemens.com

ABSTRACT

Quality assessment of speech enhancement systems has to deal with aspects such as distortion of the near-end talker's speech, and with the attenuation and distortion of the noise and the echo in different test cases. We propose first steps into the direction of a new black box objective quality assessment of speech enhancement schemes, based on our previous work on decomposition of the (enhanced) speech signal into its components *speech, (residual) noise*, and *(residual) echo*. Having these signals available, to our knowledge, for the first time a *black box* objective quality assessment of an entire speech enhancement system is proposed allowing for simultaneous measurement of, e.g., noise attenuation, echo return loss enhancement (ERLE), and perceptual evaluation of speech quality (PESQ) of the speech component in a wide range of test scenarios including double-talk. The derived scheme proves to be very useful for testing hands-free devices in practice but also for objective evaluation of sophisticated algorithms in science.

Index Terms— Objective signal quality assessment, non-blind signal decomposition, speech enhancement, hands-free

1. INTRODUCTION

In science, a comfortable way to evaluate speech enhancement algorithms is to digitally add near-end speech and noise to the echo signal and thereby construct the microphone signal. During the processing of the speech enhancement system the operational influence on the noisy microphone signal is then to be logged, and later applied individually to the speech, echo, and noise components of the microphone signal (*white box* test, e.g., [1, 2, 3]). This presumes linear processing, as can be found, e.g., in frequency domain noise reduction, where a gain is applied to the spectral amplitudes. The strength of such a method is that one achieves the three separate output signal components: The filtered speech component, the filtered echo component, and the filtered noise component, which represent the (slightly) distorted near-end talker's speech signal, the residual echo signal, and the residual noise signal, respectively.

This, however, is a highly intrusive approach, which requires access not only to the digital input and output signal of the algorithm, but also to the internal processing of the speech enhancement system. In the case of a frequency-based noise reduction, e.g., the window function, DFT frame size, and frame shift must be known, and all spectral amplitude gain values must be logged during operation. This is of course a totally impracticable test methodology if the speech enhancement system is unknown (a *black box* test is required then). It is also not useful for research on and development of more sophisticated speech enhancement systems, including also acoustic echo cancellation.

In our previous work [4], we proposed a new technique to decompose the enhanced speech signal of an unknown speech enhancement system into its three additive components speech, residual noise, and residual echo. The objective measurement results indicated that the proposed method yields a similar relative performance of (linear) noise reduction systems as the highly intrusive approach does. From informal listening tests, we also found that the mixture of the three components subjectively sounds exactly the same as the enhanced speech signal, which will be confirmed by the result of subjective listening tests presented in this paper. Moreover, we will reapply the proposed signal separation technique in (highly) nonlinear speech enhancement systems. In such systems, we will show the potential of the proposed technique to provide the additive components, whereas the highly intrusive approach might no longer be applicable due to the nonlinear processing.

This paper is organized as follows: In the next section, we briefly review the earlier proposed signal separation technique. In section 3 our choice of objective measures is discussed. Finally, section 4 presents our findings by a comparison of two exemplary speech enhancement systems using the above objective quality measures in the framework of our new *black box* test methodology.

2. ENHANCED SPEECH SIGNAL DECOMPOSITION

In this section we briefly review the steps required for signal acquisition and our previously published method of decomposing the enhanced speech signal into its components speech, noise, and echo [4].

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 assume some laboratory test setup where the near-end speech signal $s(n)$, the acoustic noise $n(n)$, and the echo $d(n)$ are digitally added to obtain the microphone signal $y(n)$.

2.1. Signal acquisition

The underlying assumption is that the microphone and the A/D converter can be modeled as a linear system, which is not too far from reality if the quantizer resolution is 16 bit or more, and if it works at a well-tuned operating point. The acquisition of the three signals $s(n), n(n)$, and $d(n)$ is accomplished as follows: First, we have to digitally record our near-end speech test signals $s(n)$ and the test

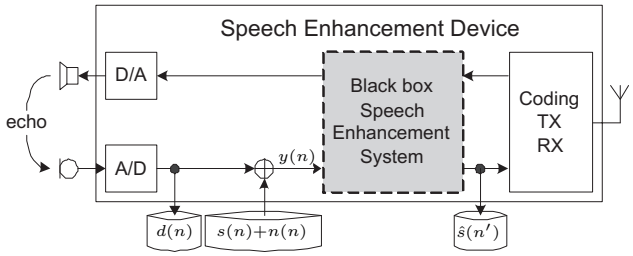

Fig. 1. Black box test of an arbitrary speech enhancement system including noise reduction and echo cancellation.

noise $n(n)$ separately via the microphone and the A/D converter of the speech enhancement device under test (see Fig. 1), and then to add both signals. In the actual real-time 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 at the microphone and digitally stored after A/D conversion. The prerecorded near-end speech plus the noise of the respective test case are then to be added in real-time to the captured echo signal $d(n)$ and are input to the black box speech enhancement system in the uplink.

Following this recording methodology, we can indeed observe the corresponding enhanced speech signal $\hat{s}(n')$ 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 $d(n), n(n), s(n), \hat{s}(n')$, with sample index n' having a certain delay with respect to sample index n. We are aware that such signals are often not yet digitally accessible in today's hands-free devices. However, they could be made accessible via a digital interface as shown in Fig. 1, as it has been proposed to ITU-T SG12 for a future speech quality assessment methodology for wideband hands-free devices. For any speech enhancement software simulation these signals should be easily available.

2.2. Signal decomposition

In a first step of the offline processing, some preprocessing with no impact on the perceived speech quality must be applied to the enhanced speech signal $\hat{s}(n')$. First, the mean of all signals should be subtracted to compensate for any DC-component. Subsequently the signal $\hat{s}(n')$ must be time aligned in accordance with the near-end speech signal $s(n)$ to obtain an enhanced speech signal $\hat{s}(n)$ with the same time index n as the input signals.

The real-time digital addition of system input components as shown in Fig. 1 is now repeated in the offline processing in the spectral domain. Spectral processing with framing, DFT, IDFT, and overlap-add is performed with the parameters N_Δ (frame shift), N (frame length), and a certain window type.

An appropriate window function and the corresponding parameters $\{N, N_\Delta\}$ have to be chosen such that the signal after frequency domain addition according to

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

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 of the frequency domain processing are then

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

Remember, in our black box test approach we are not interested in (and may not even know anything about) the internal processing of the speech enhancement system. Besides noise reduction and echo cancellation, it may perform other nonlinear processing steps. However, we simply model it 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

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

as signals $\hat{s}(n)$ and $y(n)$ are available and can be transformed into the frequency domain. Computing $G_l(k)$ directly from (3) leads to audible artifacts sounding similar to musical noise after the signal decomposition, reducing the reliability of subsequent objective and subjective measurements. This effect can be avoided by the limitation of the complex gain according to (4).

Due to linearity of the frequency domain addition in the overlap-add framework, we are now able to perform the last step of our test methodology. The filtered speech and noise components of the enhanced speech signal can be computed individually in the frequency domain by

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

and

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

Because of the limitation in (4), the sum of the filtered speech component and the filtered noise component in the frequency domain only approximates the enhanced speech signal as

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

Eqs. (5) and (6) hold for an additive mixture of filtered speech and noise components—as assumed before—but, of course, (6) holds as well for the filtered noise component only. In the latter case, the filtered echo component of the enhanced speech signal can be computed from

$$|\tilde{D}_l(k)|e^{j\phi_{\tilde{D}_l}(k)} = G_l(k) \cdot |D_l(k)|e^{j\phi_{D_l}(k)}, \tag{8}$$

and is to be added to the left-hand side of (7). The respective time domain signals $\tilde{s}(n), \tilde{n}(n)$, and $\tilde{d}(n)$ are computed by subsequent IDFT and overlap-add. They may serve for subjective listening tests—in this work, however, they are also used to compute objective quality measures.

In our previous work we found that a Blackman window of frame length $N = 512$ samples and frame shift $N_\Delta = 64$ samples yielded the best performance for 8 kHz sampled signals [4]. Thus, our experiments are performed using these settings.

3. OBJECTIVE MEASURES

For objective assessment of speech enhancement systems in this paper we choose three different widely used quantities. Note that in state-of-the-art *black box* speech quality assessment—to the best of our knowledge—it has not yet been possible to employ these quantities simultaneously before, particularly not in double-talk.

As a first measure, the amount of speech distortion is evaluated by means of PESQ-based MOS [5] of the filtered speech component $\tilde{s}(n)$ relative to the speech signal $s(n)$. PESQ scores are averaged over all test signals.

Secondly, the segmental *noise attenuation* (NA) is to be computed, defined as

$$NA_{seg} = 10\log_{10}\left[\tfrac{1}{C(\Lambda)}\sum\nolimits_{\lambda\in\Lambda} NA(\lambda)\right],$$
$$NA(l) = \frac{\sum_{\nu=0}^{N-1}\left(n^*(\nu+lN)\right)^2}{\sum_{\nu=0}^{N-1}\left(\tilde{n}^*(\nu+lN)\right)^2}. \quad (9)$$

Here, the term Λ denotes all test data frames of length N, since all are corrupted by noise. $C(\Lambda)$ is the number of frames in set Λ. Note that prior to the segmental NA computation an IIR filter is applied to the noise signal $n(n)$ and the filtered noise signal $\tilde{n}(n)$ to obtain signals $n^*(n)$ and $\tilde{n}^*(n)$, respectively. The reason for the IIR filtering is to avoid segments of a too small filtered noise signal (which may happen due to nonlinear signal processing), which may result in perceptually irrelevant outliers of the segmental NA computation.

Finally, we derive the segmental *echo return loss enhancement* (ERLE) as

$$ERLE_{seg} = \tfrac{1}{C(\Lambda_d)}\sum\nolimits_{\lambda\in\Lambda_d} ERLE(\lambda),$$
$$ERLE(l) = 10\log_{10}\left[\frac{\sum_{\nu=0}^{N-1}\left(d^*(\nu+lN)\right)^2}{\sum_{\nu=0}^{N-1}\left(\tilde{d}^*(\nu+lN)\right)^2}\right]. \quad (10)$$

Here, the term Λ_d is the test data subset with echo being present in the microphone signal. To avoid segments of a too small filtered echo signal, we likewise use the signals $d^*(n)$ and $\tilde{d}^*(n)$ computed via the same IIR filter smoothing as applied before.

4. EXPERIMENTAL RESULTS

To demonstrate the application of the signal separation method, we evaluate two speech enhancement systems consisting of noise reduction and echo cancellation sub-systems. Scheme A comprises a straightforward time-domain NLMS with VAD-controlled fixed step sizes [6, Sect. 2.2]. However, a well-performing frequency domain noise reduction based on the least square amplitude estimator [7] is used with VAD-based noise power estimation. In contrast, scheme B consists of a more sophisticated filterbank acoustic echo cancellation with near-optimum step size control [8]. Nonetheless, noise reduction is an ordinary spectral subtraction approach [9], where the noise variance is merely estimated via first-order IIR filtering with different time constants for increasing and decreasing amplitudes [10, Sect. 14.1.3]. In both schemes, a transform domain residual echo suppression is employed for further echo reduction. The actual weighting is performed in scheme B via a low-delay time domain FIR filter to reduce the signal delay.

The performance of both speech enhancement schemes with more or less a priori known strengths and weaknesses is now objectively evaluated in the new black box fashion under the NTT-AT speech and noise database. For this evaluation, a set of 48 far-end speakers (24 female and 24 male) and another set of 20 near-end speakers (10 female and 10 male) are selected and combined with 40 car noise signals. In the first experiment, $20 \times 40 = 800$ noisy near-end signals are employed to assess the *noise attenuation* performance in near-end single-talk as well as in double-talk condition. Each noisy signal is prepared for 5 SNR conditions of $0, 5, 10, 15, 20$ dB yielding $800 \times 5 = 4000$ test signals in total. For the double-talk simulations, echo signals generated from a subset of 20 far-end signals are added to the 20 near-end signals at a fixed (near-end-) speech-to-echo ratio (SER) of 5 dB prior to the addition of noise. To measure the *echo attenuation* (ERLE) performance in the noise-free double-talk case, the second experiment is performed based on $20 \times 48 = 960$ microphone signals. Here, each microphone signal is generated to obtain 3 SER conditions of $0, 5, 10$ dB resulting in $960 \times 3 = 2880$ test signals.

In the beginning, we conduct a subjective listening test to prove that, in spite of the upper limit to the gain in (4), the addition of the decomposed signals $\hat{\tilde{s}}(n) = \tilde{s}(n) + \tilde{d}(n) + \tilde{n}(n)$ is—from an auditive point of view—indistinguishable from the enhanced speech signal $\hat{s}(n)$. We randomly select 12 enhanced speech signals and their corresponding mixture signal $\hat{\tilde{s}}(n)$ comprising SNR conditions of $0, 5, 15$ dB. Next, 8 subjects have to listen to all combinations of each pair, i.e., $\left[\{\hat{s}(n), \hat{s}(n)\}, \{\hat{\tilde{s}}(n), \hat{s}(n)\}, \{\hat{s}(n), \hat{\tilde{s}}(n)\}, \{\hat{\tilde{s}}(n), \hat{\tilde{s}}(n)\}\right]$, in random sequence, and decide whether the pair of signals sounds equal or different. We subsequently compute the similarity score of all combinations from different signals, i.e., $\left[\{\hat{\tilde{s}}(n), \hat{s}(n)\}, \{\hat{s}(n), \hat{\tilde{s}}(n)\}\right]$, and that of all combinations from the same signals, i.e., $\left\{\hat{s}(n), \hat{s}(n)\}, \{\hat{\tilde{s}}(n), \hat{\tilde{s}}(n)\}\right]$. As result, we obtain similarity scores of 80.73% and 79.17% for the first and the latter cases, respectively. Consequently, listeners judge $\hat{\tilde{s}}(n)$ to sound more often similar to $\hat{s}(n)$ than two signals that are physically identical. This proves nicely, that, from an auditive point of view, our approach of signal separation yields additive components of speech, residual noise, and residual echo. It should be noted that all listeners state that they have severe problems (in) perceiving differences between the presented signal pairs at all, and consequently sometimes assume differences, where surely there were none.

After computing the error signal between $\hat{s}(n)$ and $\hat{\tilde{s}}(n)$, we also quantitatively measure the signal-to-error-signal ratio over all test signals resulting in an average value of 30.03 dB. These results further support our conclusion from the subjective listening test from above that the mixture of the component signals auditively is indistinguishable from the enhanced speech signal.

In the next experiment, we conduct objective measurements as has been explained in section 3. The results are depicted in Figs. 2 and 3. The markers of each curve in Figs. 2 and 3 represent the 5 SNR and the 3 SER conditions of the first and second experiments, respectively. Solid lines in both figures refer to double-talk simulations. The dashed lines in Fig. 2 refer to near-end single-talk cases. The more a curve is located in the upper right of the figure, the less residual noise or echo and speech distortion remain in the enhanced speech signal, and consequently the better the algorithm performs.

Analyzing the results, it can now clearly be seen that in all cases scheme A perceptually gives enhanced speech signals with less speech distortion than scheme B does. Informal listening tests in fact confirm that the enhanced speech signals produced by scheme B sound slightly metallic. From informal listening tests and our prior knowledge about the noise reduction schemes, it also turns out that scheme A with the better noise reduction approach yields less residual noise than scheme B does. This fact is clearly depicted in Fig. 2,

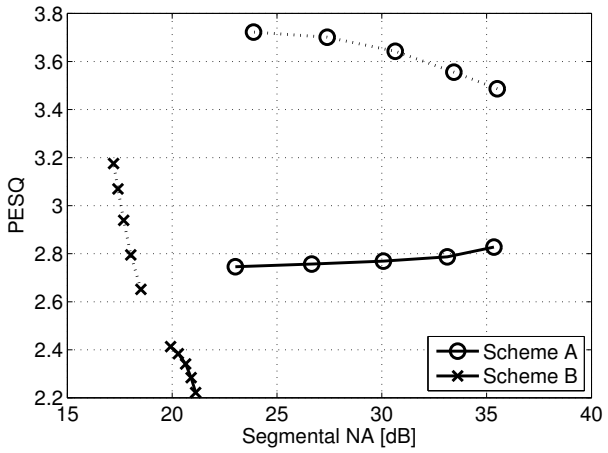

Fig. 2. Objective performance comparison of two speech enhancement systems with respect to PESQ-based MOS and segmental NA in near-end single-talk (dashed) and double-talk (solid) condition.

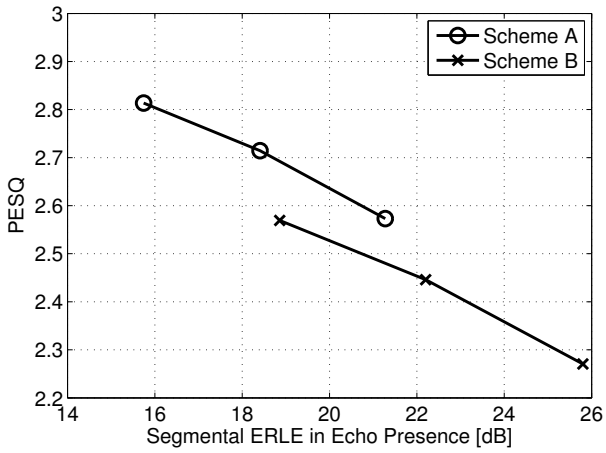

Fig. 3. Objective performance comparison of two speech enhancement systems with respect to PESQ-based MOS and segmental ERLE in double-talk condition.

where the near-end single-talk and double-talk curves of scheme A are located to the right of the respective curves of scheme B.

To analyze the echo attenuation performance, we compare the performance in far-end single-talk and double-talk cases. The far-end single-talk performance (SER $\rightarrow -\infty$ dB) is measured by applying only echo signals generated from all 48 far-end speaker signals to the speech enhancement systems, and it results in $ERLE_{\text{seg}} = 33.33$ dB for scheme A, and $ERLE_{\text{seg}} = 51.76$ dB for scheme B. Along with Fig. 3, we can easily see that scheme B with a better echo cancellation technique leads to a higher echo attenuation (ERLE) than scheme A does, which is to be expected. Nevertheless, similar to the result of the previous experiment, scheme B unfortunately still shows more perceptual speech distortion.

Based on these results, we have shown that our signal separation technique along with standard measures such as PESQ, ERLE, and noise attenuation can provide a powerful tool to simultaneously assess aspects of speech quality, that previously were available only in different test cases with highly specialized test sequences, and that was not possible to be measured in double-talk before.

5. CONCLUSIONS

In this paper we show how to report on objective speech quality measures such as PESQ, ERLE, and noise attenuation for speech enhancement systems in a black box test scenario. Our approach allows the *simultaneous* measurement of all 3 quantities *even in double-talk* test conditions. As a prerequisite, we use a simple but effective scheme to decompose the enhanced speech signal into its three components (distorted) speech, residual noise, and residual echo. Comparing two speech enhancement systems, weaknesses and strengths w.r.t. noise reduction and acoustic echo cancellation are clearly reported—also in double-talk situations. This new methodology provides engineers and scientists with a powerful means to measure the performance of hardware hands-free systems or of algorithm simulations, respectively, in a black box type of test.

6. REFERENCES

[1] T. Lotter, *Single and Multimicrophone Speech Enhancement for Hearing Aids*, Ph.D. thesis, Aachener Beiträge zu digitalen Nachrichtensystemen, edited by P. Vary, vol. 9, (ISBN 3-86073-438-5), 2004.

[2] T. Fingscheidt and S. Suhadi, "Data-Driven Speech Enhancement," in *Proc. of ITG-Fachtagung "Sprachkommunikation"*, Kiel, Germany, Apr. 2006, VDE–Verlag.

[3] S. Suhadi, S. Stan, and T. Fingscheidt, "A Novel Environment-Dependent Speech Enhancement Method with Optimized Memory Footprint," in *Proc. of International Conference Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006.

[4] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in *Proc. of INTERSPEECH'07*, Antwerp, Belgium, Aug. 2007.

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

[6] 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.

[7] C. Beaugeant and P. Scalart, "Speech Enhancement Using a Minimum Least Square Amplitude Estimator," in *Proc. of International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, Sept. 2001, pp. 191–194.

[8] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Low-Delay Subband Acoustic Echo Control in an Automotive Environment," in *Proc. of Biennial on DSP for In-Vehicle and Mobile Systems*, Istanbul, Turkey, June 2007.

[9] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[10] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley-Interscience, Hoboken, N.J., 2004.