HANDS-FREE SYSTEM WITH LOW-DELAY SUBBAND ACOUSTIC ECHO CONTROL AND NOISE REDUCTION

Kai Steinert[1], Martin Schönle[1], Christophe Beaugeant[2], and Tim Fingscheidt[3]

[1]Siemens AG, Corporate Technology, Munich, Germany
[2]Infineon Technologies, Business Group Communication Solutions, Sophia-Antipolis, France
[3]Institute for Communications Technology, Braunschweig Technical University, Germany
E-Mail: {kai.steinert.ext,martin.schoenle}@siemens.com, christophe.beaugeant@infineon.com, t.fingscheidt@tu-bs.de

ABSTRACT

Echo cancellation and noise reduction for hands-free systems are challenging tasks in speech signal processing. The presence of strong local speech and noise and a changing acoustical enclosing may severely impair the performance of the algorithms. Usually additional constraints such as a low signal delay are also requested for real time implementation.

We present a hands-free system consisting of a delayless subband adaptive filter with a low-delay echo and noise suppression postfilter. All parameters are estimated in the subband domain, whereas the filtering takes place in the time domain. Thus, our system has a significantly lower processing delay than similar proposals. We compare its performance with respect to echo and noise attenuation and speech distortion with a state-of-the-art hands-free system in a simulated car environment.

Index Terms— Hands-free system, echo cancellation, speech enhancement, low-delay filter, subband system

1. INTRODUCTION

We consider acoustic echo cancellation and noise reduction for hands-free applications such as telecommunication. Usually an adaptive filter is employed to estimate the echo path between loudspeaker and microphone, and thereby create an echo signal estimate [1]. After its subtraction from the microphone signal, typically a residual echo signal remains, due to the limited adaptive filter order, nonlinearities and time-variability in the echo path, and local speech and noise activity. The residual echo signal—as well as the local noise—is attenuated by weighting in a transform (e.g., FFT) domain [1].

Subband implementations of adaptive filters offer a reduced computational complexity, an increased convergence speed, and additional degrees of freedom (e.g., a different filter length in each band) [1]. Their drawback, however, is a signal delay that is due to the analysis and synthesis filtering process. For the same user acceptance of an additional delay in the system, a higher echo attenuation has to be provided [1, sec. 3]. Furthermore tight constraints of conversational applications, e.g., 30 ms in the uplink path of a car-kit hands-free system according to [2], necessitate a low-delay system. In a delayless subband adaptive filter structure [3] the filter coefficients are adapted in the subband domain, transformed to a broadband FIR impulse response, and the actual filtering is carried out on a sample-by-sample basis in the time domain with virtually no further delay. However, filtering in the fullband rather than in the subbands again adds some complexity and cannot prevent a delay in coefficient adaptation and tracking taking place in the subbands.

Likewise, if the residual echo and noise suppression postfilter weights are calculated in the subband domain as well, a broadband FIR filter of low degree can be approximated for time domain filtering [4]. This way a delay much smaller than that caused by the filter bank analysis stage is feasible.

As novelty, we present a hands-free system in which a delayless subband adaptive filter is combined with a low-delay postfilter for residual echo suppression and noise reduction. Thereby we are able to achieve a total signal delay much smaller than half of that caused by the same system with filtering and weighting in the subband domain. We compare our system with a state-of-the-art hands-free reference system [5] in a simulated car environment. In particular we focus on the echo and noise attenuation and the speech distortion. An experiment evaluating the tracking performance is considered as well.

This paper is organized as follows. The next section presents the system to be discussed in detail. We describe its components and their control, that is, the coefficient calculation and application. In section 3 we describe our measurement setup and the objective measures used. We present our findings in section 4 and conclude with a discussion thereof.

2. SYSTEM DESCRIPTION

The echo cancellation and noise reduction system to be discussed is shown in Fig. 1. The microphone signal $y(n)$ is composed of the local (useful) speech signal $s(n)$, the local noise signal $n(n)$, and the echo $d(n)$ generated by the far-end (loudspeaker) signal $x(n)$. Both the microphone and the loudspeaker signal are decomposed into the subband signals $Y_k(m)$ and $X_k(m)$, respectively, where k indicates the subband and m the subband time index. In each subband k an adaptive filter calculates N_k subband filter coefficients $\hat{\mathbf{H}}_k(m) = [\hat{H}_{k,0}(m), ..., \hat{H}_{k,N_k-1}(m)]^T$. Thereby the subband impulse responses are estimated that correspond to the time domain loudspeaker-enclosure-microphone (LEM) system $\mathbf{h}(n) = [h_0(n), ..., h_{N-1}(n)]^T$ assumed to be of a certain length N. From the estimated subband coefficients, a time domain FIR filter $\hat{\mathbf{h}}(n) = [\hat{h}_0(n), ..., \hat{h}_{N-1}(n)]^T$ (same length N) for delayless fullband filtering is obtained via a weight transform [3]. Similarly, residual echo suppression and noise reduction postfilter weights $G_k(m)$ are calculated in the subbands and transformed to a low-delay time domain moving-average filter $\mathbf{g}(n)$ [4]. Hence, postfiltering has a lower delay than an analysis-synthesis filter bank system. The output signal is therefore calculated as

$$\hat{s}(n) = \big(y(n) - x(n) * \hat{\mathbf{h}}(n)\big) * \mathbf{g}(n). \quad (1)$$

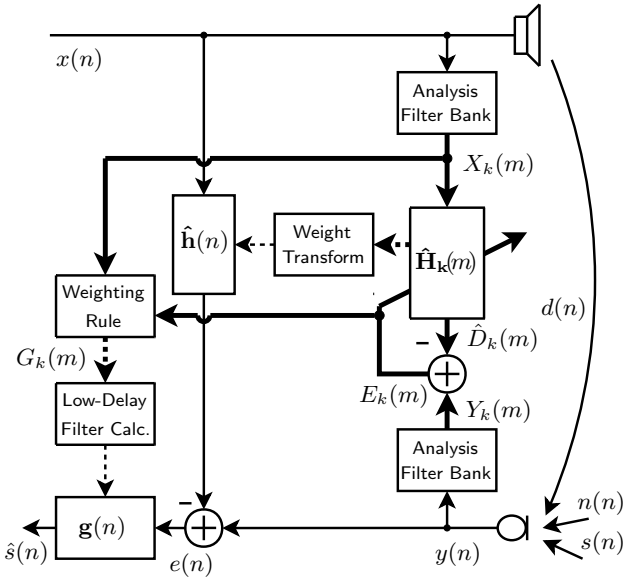

Fig. 1. Proposed low-delay subband system. Thin lines stand for time domain, thick lines for subband domain signal paths, and dashed lines indicate a coefficient or weight transfer.

2.1. Subband decomposition

To ensure a sufficiently good performance of the adaptive filter without the need of cross-band adaptation [1, app. B], we employ a uniform DFT-modulated polyphase filter bank. We chose 32 channels at a subsampling rate of 16 ("Analysis Filter Bank" blocks in Fig. 1). Our (power complementary) prototype filters are designed according to [1, app. B] with a length of 128 coefficients and a stopband attenuation of approximately 72 dB.

2.2. Delayless subband adaptive filter

We have implemented a subband NLMS adaptive filter [1, chap. 7] (block "$\hat{\mathbf{H}}_\mathbf{k}(m)$" in Fig. 1) with the near-optimum step size formula according to [1, chap. 13]. We use an open-loop adaptation scheme [3] described by the subband filtering, the step size calculation, and the subband filter adaptation

$$E_k(m) = Y_k(m) - \hat{\mathbf{H}}_k^H(m)\mathbf{X}_k(m), \tag{2}$$

$$\mu_k(m) = \frac{E\{|E_{u,k}(m)|^2\}}{E\{|E_k(m)|^2\}}, \text{ and} \tag{3}$$

$$\hat{H}_k(m+1) = \hat{H}_k(m) + \mu_k(m)\frac{\mathbf{X}_k(m)E_k^*(m)}{||\mathbf{X}_k(m)||}, \tag{4}$$

respectively, with the subband excitation signal vector $\mathbf{X}_k(m) = [X_k(m), X_k(m-1), ..., X_k(m-N_k+1)]^T$, the step size $\mu_k(m)$, and the undisturbed error $E_{u,k}(m) = E_k(m) - S_k(m) - N_k(m)$ with $S_k(m)$ and $N_k(m)$ the subband representation of the local speech and noise signal, respectively. The expression $E\{\cdot\}$ stands for the expectation value, $|\cdot|$ is the magnitude, $||\cdot||$ the Euclidean vector norm, $(\cdot)^*$ the complex conjugate, and $(\cdot)^H$ the Hermitian vector. The estimation of the undisturbed error power will be addressed below. To save memory, the subband-dependent adaptive filter length N_k may be chosen shorter for higher-frequency subbands where the impulse response of certain rooms decays faster [1, chap. 9]. We chose filter lengths between 5 (highest bands) and 16 (second lowest band).

Since the band filters providing $X_k(m)$ cause a time delay, the optimum subband impulse response is non-causal [1, chap. 9]. These non-causal taps could be modeled by (artificially) introducing a delay into the echo path, thus, however, increasing the round-trip delay. To keep the signal delay low, we do not apply this method.

Finally, a weight transform [3] is used to calculate the fullband FIR filter $\hat{\mathbf{h}}(n)$ from the subband coefficients $\hat{\mathbf{H}}_\mathbf{k}(m)$.

2.3. Low-delay postfilter

The postfilter comprises a Wiener filter weighting rule (as proposed in [6]) for residual echo suppression and one for noise reduction. By separating both tasks into individual filters, a flexible system can be realized so that, e.g., the residual echo suppression may easily be deactivated. The weights $G_k^p(m)$ of each filter are calculated according to

$$G_k^p(m) = \frac{\xi_k^p(m)}{1+\xi_k^p(m)} \tag{5}$$

where $p \in \{e_u, n\}$ denotes the type of perturbation, that is, residual echo or noise, respectively, and

$$\xi_k^p(m) = \frac{E\{|S_k^p(m)|^2\}}{E\{|P_k(m)|^2\}} \tag{6}$$

the a priori SPR (signal-to-perturbation ratio). $S_k^p(m)$ is the useful signal with respect to the type of perturbation to be attenuated, i.e., $S_k^{e_u}(m) = S_k(m) + N_k(m)$ and $S_k^n(m) = S_k(m) + E_{u,k}(m)$. Here $P_k(m)$ is the perturbation signal, residual echo $E_{u,k}(m)$ or noise $N_k(m)$, respectively. The estimate of $\xi_k^p(m)$, $\hat{\xi}_k^p(m)$, is calculated with a decision-directed approach as follows

$$\hat{\xi}_k^p(m) = \alpha_p \frac{|\hat{S}_k^p(m-1)|^2}{E\{|P_k(m-1)|^2\}} + (1-\alpha_p)\max\{\gamma_k^p(m)-1, 0\} \tag{7}$$

with $0 < \alpha_p < 1$, $\gamma_k^p(m)$ is the a posteriori SPR defined as

$$\gamma_k^p(m) = \frac{|E_k(m)|^2}{E\{|P_k(m)|^2\}}. \tag{8}$$

The instantaneous local signal power of the last iteration is calculated according to $|\hat{S}_k^p(m-1)|^2 = \left(G_k^p(m-1)\right)^2 \cdot |E_k(m-1)|^2$. The estimation of the residual echo power $E\{|E_{u,k}(m)|^2\}$ and the noise power $E\{|N_k(m)|^2\}$ will be the subject of the following subsection.

The global weights $G_k(m) = G_k^{e_u}(m) \cdot G_k^n(m)$ are transformed into a time domain FIR postfilter $\mathbf{g}'(n)$, cf. block "Low-Delay Filter Calc." in Fig. 1 [4]. This filter mathematically corresponds to an analysis-synthesis subband system with application of the spectral weights $G_k(m)$ in the subbands. Unlike the filter bank discussed in section 2.1 and in order to reduce the signal delay, subsampling does not take place and the synthesis filter bank stage consists of a mere summation of the subbands without another filtering. The analysis prototype filter is an M-th band filter (amplitude complementary) of length 128. A further reduction of the signal delay is achieved by truncating $\mathbf{g}'(n)$ to a length of 32 samples to obtain $\mathbf{g}(n)$ (cf. block "$\mathbf{g}(n)$" in Fig. 1).

Similar to the case of the delayless subband adaptive filter, the postfilter coefficient computation is delayed relative to its application in the time domain. However, for our system we were not able to perceive a degraded signal quality due to this condition.

SNR	-5	0	5	10	15	20
ERLE Ref.	16.7	18.2	19.0	19.4	19.9	20.8
ERLE Proposal	20.3	21.5	22.3	22.3	22.2	21.9
NA Ref.	16.4	16.6	15.7	13.8	12.0	10.4
NA Proposal	16.1	16.1	16.0	15.7	15.2	14.6

Table 1. ERLE and NA results for both systems for far-end single-talk at different SNRs. All values are given in dB.

2.4. Control of the adaptive filter and postfilter

Both the adaptive filter and the residual echo suppression postfilter require an estimate of the undisturbed echo power $E\{|E_{u,k}(m)|^2\}$. It can approximately be written as [1, chap. 13]

$$E\{|E_{u,k}(m)|^2\} \approx E\{|X_k(m)|^2\} \cdot \beta_k(m) \tag{9}$$

where $\beta_k(m) = E\{||\mathbf{H}'_k(m) - \hat{\mathbf{H}}'_k(m)||^2\}$ stands for the so-called system distance. The true echo path impulse response $\mathbf{H}'_k(m)$ and the adaptive filter impulse response $\hat{\mathbf{H}}'_k(m)$ are defined similarly to $\hat{\mathbf{H}}_k(m)$, but additionally contain the non-causal taps. We use an approach based on the cross spectral method to calculate $\beta_k(m)$. Basically, in each subband the system distance is calculated as the quotient of a cross correlation and an auto correlation term. For a detailed explanation, the reader is referred to [7]. That way, an additional echo path change detector or an explicit model of impulse response variations is not needed.

The noise power is estimated in each subband by first order IIR filter smoothing with large smoothing constants for rising signal edges and small smoothing constants for falling edges [1, chap. 14]. As a result, the minimum of the noisy signal is being tracked.

3. SIMULATION SETUP AND INSTRUMENTAL MEASURES

We compared the performance of our proposal with a reference state-of-the-art hands-free system [5]. The latter consists of a time domain NLMS echo canceller with VAD-controlled fixed step sizes and a frequency domain residual echo and noise suppression. The noise estimation is based on a three-step VAD. The gain loss control has been deactivated.

We used four male and four female American English speakers for far-end and local speech and four car noise signals, all at 8 kHz. The SNR (local-speech-to-local-noise ratio) was varied between -5 and 20 dB in 5 dB steps. The echo signal was created by convolving the excitation with a car impulse response of length 256 and adding the result to the local speech and noise with an SER (local-speech-to-echo ratio) of 0 and 5 dB.

As a measure of the echo attenuation, we consider the echo return loss enhancement (ERLE)

$$ERLE = \frac{1}{C(N_d)} \sum_{n \in N_d} 10 \log_{10} \left(\frac{E\{d^2(n)\}}{E\{\tilde{e}_u^2(n)\}} \right) \tag{10}$$

as an average over the set N_d of all $C(N_d)$ samples of echo presence of all signals with $\tilde{e}_u(n)$ the weighted residual echo signal (i.e., the residual echo part of the enhanced output signal). For the tracking experiment a temporal ERLE was calculated according to

$$ERLE_t(n) = 10 \log_{10} \left(\frac{E\{d^2(n)\}}{E\{\tilde{e}_u^2(n)\}} \right). \tag{11}$$

The noise attenuation

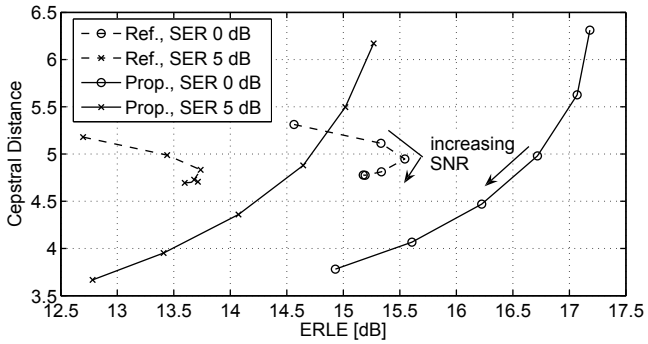

Fig. 2. Cepstral distance and ERLE for the reference (dashed lines) and the proposed system (solid lines) for different SERs (0 dB for circular markers, 5 dB for cross markers), and for different SNRs (markers) during double-talk condition.

$$NA = 10 \log_{10} \left(\frac{1}{C(N_n)} \sum_{n \in N_n} \frac{E\{n^2(n)\}}{E\{\tilde{n}^2(n)\}} \right) \tag{12}$$

is evaluated over all samples $C(N_n)$ of the set N_n (noise present, i.e., the entire signal) of all signals. Thereby $\tilde{n}(n)$ is the weighted noise signal (i.e., the residual noise part of the enhanced output signal). The speech distortion in case of double-talk was measured by the cepstral distance *CD* calculated as follows.

$$C_s^{(\lambda)}(i) = \mathrm{IDFT}\{\ln|\mathrm{DFT}\{S^{(\lambda)}(i)\}|\}, i = 0, ..., N-1 \tag{13}$$

$$CD^{(\lambda)} = \frac{\sqrt{[C_s^{(\lambda)}(0) - C_{\tilde{s}}^{(\lambda)}(0)]^2 + 2\sum_{i=1}^{N_{CD}-1}[C_s^{(\lambda)}(i) - C_{\tilde{s}}^{(\lambda)}(i)]^2}}{\ln(10)} \tag{14}$$

$$CD = \frac{1}{C(M_s)} \sum_{\lambda \in M_s} CD^{(\lambda)}. \tag{15}$$

Here $C_s^{(\lambda)}(i)$ and $C_{\tilde{s}}^{(\lambda)}(i)$ denote a cepstral coefficient of the speech and the weighted speech signal (i.e., the speech part of the enhanced output signal), respectively, and $S^{(\lambda)}(i) = s(\lambda \cdot l + i)$ with the frame shift $l \in \mathbb{N}$. Similar to the averaged ERLE, $C(M_s)$ is the number of frames of the set M_s with local speech being present.

The weighted residual echo, weighted noise, and weighted speech signals were obtained from the enhanced output signal and the clean input signals using the signal separation technique [8].

4. SIMULATION RESULTS

In a single-talk experiment, we chose two utterances for each speaker (i.e., 16 utterances altogether) as far-end excitation. The near-end signal only consists of noise of the SNRs mentioned in section 3. The results in terms of the averaged ERLE and NA values are given in Tab. 1. It can be seen that, for this experiment, the proposed system exhibits an ERLE that is increased by a few decibels compared to the reference system, particularly for low SNRs. This can be expected due to transform domain filtering and the near-optimum step size that implicitly takes into account the local noise level. The NA of the new system is slightly worse for low SNRs, but is improved for higher SNRs. We found that the residual noise of the proposed system sounds somewhat more similar to the clean noise with respect to the spectral coloration, whereas the residual noise of the reference

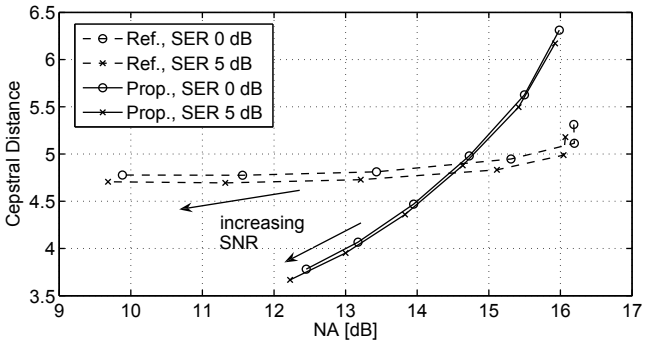

Fig. 3. Cepstral distance and noise attenuation for the reference (dashed lines) and the proposed system (solid lines) for different SERs (0 dB for circle markers, 5 dB for cross markers) and for different SNRs (markers) during double-talk condition.

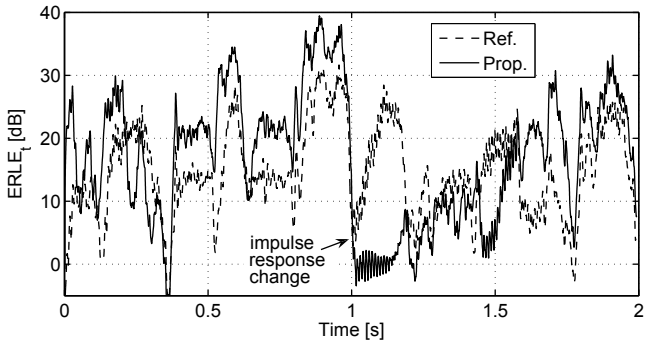

Fig. 4. Tracking performance for male speech excitation and white background noise of 15 dB SNR.

system sounds more white. Yet the output signal of the proposal contains short instants of stronger residual noise as the noise estimation cannot track fast increases in noise power.

To assess the speech distortion caused by the system, we conducted double-talk experiments. We considered the far-end/near-end combination of the eight speakers m/m, m/f, f/f, f/m with four utterances each. The SNRs and SERs were chosen as given in Sec. 3.

The ERLE result is depicted in Fig. 2. As is expected, we obtain clearly different ERLE values for different SERs. CD and ERLE do not vary much for the reference system when the SNR is changed. There is a much larger variance for the proposed system. Specifically, for low SNRs, our system can achieve a higher echo attenuation, but, at the cost of a larger signal distortion. For higher SNRs the ERLE as well as the CD tend to be smaller compared to the reference system. The ERLE improvement for low SNRs can be explained by the near-optimum step size and the continuous adaptation in the subband structure without an explicit VAD.

Apparently, the residual echo of our system is concentrated in lower-frequency areas, whereas that of the reference system is more equally distributed over the frequency range. For the low-SNR cases, parts of the enhanced output signal may sound slightly robot-like for the proposal. However, for cases of higher SNRs, we perceived the output signal to be more similar to the clean signal than for the reference system.

The result in terms of the NA in relation to the CD is presented in Fig. 3. Obviously the influence of the SER on both the CD and the NA is quite small. For high SNRs, the proposed system yields a higher NA at a smaller speech distortion. For lower SNRs approximately the same NA is obtained for both systems, with the presented system having a larger CD. The residual noise of the proposed system tends to be weaker during local speech absence, but is stronglier modulated by the local speech than for the reference system. The performance of the system could be improved with a higher spectral resolution at the expense of a larger adaptation delay.

The tracking performance was evaluated with male speech in a single-talk situation with white background noise of 15 dB SNR. It can be seen (Fig. 4) that the reference system reconverges faster which can be explained by the delayed subband tracking with respect to the time domain filtering in the delayless adaptive filter. Yet in general, the proposed system can achieve a higher ERLE.

5. CONCLUSION

We have proposed a low-delay hands-free system comprising echo cancellation and noise reduction which is based on a delayless subband adaptive filter with time domain FIR postfiltering. The algorithm does not require an explicit VAD or double-talk detector. We evaluated the performance in comparison to a state-of-the-art reference hands-free system for a simulated car environment. The results indicate that our system may lead to a far better echo and noise attenuation. However, the only part to be improved is the performance for drastically high noise levels. Our system achieves a signal delay of around 16 samples compared to 127 samples for an analysis-synthesis system with a similar filter bank and 40 samples for the reference system. The presented system is an attractive alternative for applications where low signal delay is essential such as hands-free telecommunication.

6. REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*, Wiley, Hoboken NJ, 2004.

[2] Verband der Automobilindustrie, "VDA specification for car hands-free terminals," Version 1.5, Draft, Dec. 2004.

[3] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Signal Processing Mag.*, vol. 43, no. 8, pp. 1819–1830, Aug. 1995.

[4] H. W. Löllmann and P. Vary, "Uniform and warped low delay filter-banks for speech enhancement," *Speech Communication (Elsevier)*, vol. 49, pp. 574–587, 2007.

[5] 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. AES, 29th conference*, Seoul, South Korea, Sept. 2006.

[6] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication (Elsevier)*, vol. 49, pp. 588–601, 2007.

[7] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Low-delay subband echo control in an automotive environment," in *Proc. Biennial on DSP for in-Vehicle and Mobile Systems*, Istanbul, Turkey, June 2007.

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