Chapter 6
A Wideband Automotive Hands-Free System for Mobile HD Voice Services

Marc-André Jung and Tim Fingscheidt

Abstract Wideband mobile telephony supporting a speech bandwidth from 50 to 7,000 Hz gets more and more employed. These so-called mobile HD Voice services consequently find their way into automobile applications. In this chapter we present a wideband hands-free system for automotive telephony applications with a synchronously adapted acoustic echo canceller and postfilter. It is based on a frequency domain adaptive filter approach and Kalman filter theory and makes use of a generalized Wiener postfilter for residual echo suppression and noise reduction in a consistent way. To provide a high convergence rate in case of time-variant echo paths, the echo canceller with very robust double-talk performance is supported by a fast converging shadow filter, which allows for a good tracking performance. A decimation approach is used to decrease algorithmic delay and computational complexity without loss of quality. Experimental results with car cabin impulse responses show good echo cancellation capabilities with fast convergence times along with extraordinary full-duplex performance while still keeping an almost untouched speech component in the converged state.

Keywords AEC • FDAF • Hands-free system • HD voice • Shadow filter • Wideband

6.1 Introduction

Mobile HD Voice services supporting wideband speech (50–7,000 Hz) as opposed to narrowband speech (300–3,400 Hz) allow for a high-quality and high-intelligibility telephony experience. Syllable articulation (i.e., human syllable recognition rate) increases from 90 % to about 98 %, making the use of spelling alphabets for

M.-A. Jung (✉) • T. Fingscheidt
Institute for Communications Technology, Technische Universität Braunschweig,
Schleinitzstr. 22, Braunschweig D-38106, Germany
e-mail: jung@ifn.ing.tu-bs.de; fingscheidt@ifn.ing.tu-bs.de

G. Schmidt et al. (eds.), *Smart Mobile In-Vehicle Systems: Next Generation Advancements*, DOI 10.1007/978-1-4614-9120-0_6,

proper names widely obsolete. Apart from proper names, in narrowband speech transmission listeners typically are able to employ their language model in order to reconstruct missing syllables in an interpolative fashion. This, however, does not work sufficiently well in the case of foreign-language conversations. Also in situations with a high level of background noise—as it is typical for mobile telephony in general and automotive telephony in particular—the further drop in syllable articulation cannot be compensated sufficiently for. All these aspects were driving forces for the worldwide deployment of mobile wideband speech services in the past years, commonly being known as mobile HD Voice services.

High-quality hands-free capabilities are a greatly demanded feature of telecommunication systems in office, home, or car environments and—referring to the latter—are even mandatory in many countries. Several state-of-the-art algorithms have been developed to fulfill technical requirements, such as full-duplex speech transmission capability, sufficient acoustic echo cancellation even for highly time-variant echo paths, and minimal speech distortion (e.g., [1–6]). Nevertheless, those requirements often collide with practical restrictions such as low complexity and algorithmic delay [7–9].

Hands-free systems are usually designed to cope with signal degradations stemming from the acoustic environment. These degradations are typically caused by acoustic echo and additive noise, leading to reduced intelligibility and speech quality. This is specifically the case for long round-trip delays or high noise immissions, as can be often found in automotive mobile phone usage. As a countermeasure, acoustic echo cancellers (AECs) [1, 6, 10, 11] and postfilters (PFs) for residual echo suppression (RES) [12, 13] and noise reduction (NR) approaches [14–16] have been proposed, typically working at a sampling rate of $f_s = 8$ kHz (narrowband speech).

With upcoming mobile wideband speech transmission (HD Voice services) at a sampling rate of $f_s = 16$ kHz, there are a couple of obstacles to be solved when designing a hands-free system. The doubled sampling rate causes a non-negligible increase of algorithmic complexity and can also lead to other unwanted effects when porting an algorithm from narrowband to wideband [9].

Typical hands-free system representatives in the time domain are based on the normalized least mean square (NLMS) [17], affine projection (AP) [17–19], recursive least squares (RLS) [20], or Kalman algorithm [6, 21]. These approaches usually feature a simple algorithmic structure with the ability to work on a per-sample base. On the one hand, this usually leads to zero or low delay; on the other hand, modeling of longer impulse responses (IRs) can lead to exceedingly high computational complexity if the filter is adapted in every single sample. This problem can be addressed by block processing, where the filter is only adapted once per block of samples. Albeit computationally efficient, this block processing leads to algorithmic delay and a slower convergence rate. Due to the fact that most of these algorithms make the assumption of a spectrally white echo signal but speech signals usually still have some inherent correlation, adaptation can only take place in the limited direction of the error signal vector. This decreased convergence rate can partly be avoided by using some kind of decorrelation technique for the excitation signal [6]. Whereas convergence speed can be increased especially with the RLS and Kalman algorithms, tracking performance often still suffers since adaptation of a well-converged system model to IR changes only takes place in little steps [7]. Another well-known problem of time domain AEC approaches is the poor double-talk performance. Presence of near-end speech or noise leads to undesired adaptation and therefore misestimation of the true impulse response. To avoid this, a—more or less—robust double-talk detection (DTD) scheme is often applied [7], which triggers an adaptation speed reduction during double-talk.

Adaptation in a transform domain like subband or frequency domain may circumvent some of the abovementioned deficiencies. However, it should be mentioned that transformation domain processing may introduce other, possibly more perturbing, problems. Having said this, these algorithms may be a very good choice if applied appropriately. Filter adaptation in the subband domain, for example, can lead to a significantly reduced computational complexity if long impulse responses have to be modeled. This is made possible by splitting the fullband signal into several subbands by means of a filter bank. Due to this, each of the subband signals is analyzed separately, whereas subsampling can be applied, and individual filter lengths for each subband can be chosen. Furthermore, convergence speed is highly improved since each subband signal can be esteemed as spectrally white. It should be considered, however, that algorithmic delay increases. Note that also low-delay filter-bank approaches exist, e.g., [4, 22]. However, the problem of poor double-talk performance with the need of DTD often remains. Furthermore, the design of a filter-bank analysis and synthesis structure is typically realized with prototype filters [23], which might be tedious to some extent.

As an alternative to these subband algorithms, the so-called frequency domain adaptive filter (FDAF) algorithms can be used [24]. Adaptation of the impulse response model and estimation of the echo signal is performed in the frequency domain. This allows to compute frequency-dependent parameters like optimal stepsize vectors. In our case, the inversely transformed estimated echo signal is then used to filter the microphone signal in the time domain [3, 25]. Due to the inherent block processing of the fast convolution in the frequency domain, in many cases complexity can be drastically decreased. A further significant advantage is the extraordinary double-talk performance of some FDAF algorithms, which makes DTD obsolete [3, 25, 26]. Furthermore, they are able to preserve a very good quality of the speech component in the uplink (send) path. Unfortunately, having to buffer a block of samples for the discrete Fourier transform (DFT) introduces delay in the uplink signal path. As another drawback, large DFTs, as they are needed to sufficiently cover long impulse responses, also lead to comparably slow convergence times.

Since AEC filters typically achieve a yet insufficient amount of echo suppression, a subsequent postfilter is needed. This also covers nonlinear echo components and can additionally serve as NR filter [12]. Whereas time domain AEC algorithms are frequently complemented with time domain gain loss control (GLC) postfilters [5, 7, 27], transform domain AEC filters often make use of postfilters within the same domain [13, 22, 26, 28, 29]. Here, the group of GLC postfilters could be shortly described as computationally efficient with the drawback of poor double-talk performance, while the group of frequency domain or subband postfilters often show better performance—especially during double-talk—with the drawback of additional signal delay.

The focus of our work lies in a wideband hands-free system for automotive applications with relatively short impulse responses. In contrast to the mobile use case, here the demand of very low complexity is of subordinate importance. A well-balanced double-talk performance, on the contrary, is a crucial point to keep the mental distraction of the driver at a low level. Additionally, algorithmic delay should be kept low to avoid a large contribution to the round-trip delay. Due to its excellent double-talk performance with still tolerable algorithmic delay, an FDAF-based Kalman filter algorithm [3, 25] is chosen for the following investigations and implemented for wideband speech. The algorithm is supplemented with a shadow filter (SF), which leads to a drastic reduction of convergence time. Furthermore, a modified postfilter setup is suggested, which is able to significantly reduce algorithmic delay at a given echo suppression by means of decimation in the DFT domain.

In Sect. 6.2 the FDAF hands-free algorithm based on [3, 25] is presented but already adopted to wideband speech. Section 6.3 presents the latency reduction by decimation in the DFT domain as well as the shadow-filter-enhanced FDAF algorithm. In Sect. 6.4 experimental results of single- and double-talk simulations are given. Echo suppression, convergence behavior, algorithmic delay, and quality of the speech component are discussed.

6.2 State-of-the-Art FDAF

We have motivated before that hands-free systems with adaptation of the filter coefficients in the frequency domain are generally a good choice if low computational complexity for long impulse responses, good double-talk performance, and little degradation of the near-end speech signal component are desired. The adaptive filter is placed in parallel to the electroacoustic echo path or loudspeaker-enclosure-microphone (LEM) system, trying to estimate a replica echo signal. In case of the FDAF algorithm, the adaptation of the filter coefficients and the computation of the estimated echo signal are performed in the frequency domain.

As depicted in Fig. 6.1, in a digital model of the LEM system, the echo signal $d(n)$ with sample index n is the result of the convolution of the far-end signal $x(n)$ with the LEM impulse response. The microphone signal is then given by $\boldsymbol{y} = [y(n-R+1), \ldots, y(n)]^T$, with R being the frame shift, also called block length, $[\cdot]^T$ being the transpose, and $y(n) = s(n) + n(n) + d(n)$, whereas $s(n)$ is the near-end speech signal and $n(n)$ is the noise component.

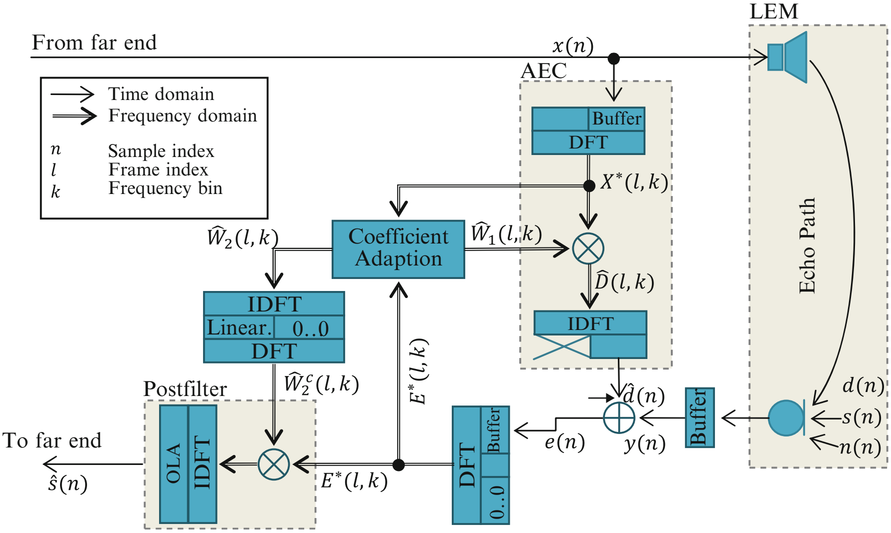

Fig. 6.1 State-of-the-art FDAF-based hands-free system

Then the loudspeaker signal $x(n)$ is transformed into the DFT domain by

$$\begin{aligned}\boldsymbol{X}_l &= [X(l,0), \ldots X(l,k), \ldots X(l,K-1)]^T \\ &= DFT\left\{[x(n-K+1), \ldots, x(n-R), x(n-R+1), \ldots, x(n)]^T\right\},\end{aligned}$$

with frame index l and frequency bin index k. By making use of the FDAF approach based on Kalman filter theory, the DFT domain adaptive filter coefficients $\hat{W}_1(l,k)$ are estimated [26, 28]. An estimate of the frequency domain replica echo signal is then computed by

$$\hat{D}(l,k) = \hat{W}_1(l,k) \cdot X^*(l,k) \tag{6.1}$$

for $k = 0, \ldots, K-1$, with $(\cdot)^*$ being the conjugate complex operator. Its inverse DFT (IDFT) delivers $\left[\ldots, \hat{\boldsymbol{d}}_l^T\right]^T = IDFT\{\hat{\boldsymbol{D}}_l\}$, with $\hat{\boldsymbol{D}}_l = \left[\hat{D}(l,0), \ldots, \hat{D}(l,K-1)\right]^T$ and $\hat{\boldsymbol{d}}_l = \left[\hat{d}(n-R+1), \ldots, \hat{d}(n)\right]^T$, which is then used to compute R samples of an error signal

$$e(n) = y(n) - \hat{d}(n). \tag{6.2}$$

The residual echo $r(n) = d(n) - \hat{d}(n)$ is contained in the error signal as $e(n) = r(n) + s(n) + n(n)$. The DFT error signal

$$\boldsymbol{E}_l = DFT\left\{\left[\boldsymbol{0}^T_{K-R-O}, \left(\boldsymbol{e}^-_{l-1}\right)^T, \boldsymbol{e}^T_l\right]^T\right\} \tag{6.3}$$

with $\boldsymbol{0}_{K-R-O}$ being a (K-R-O)-dimensional zero vector, $\boldsymbol{e}^-_{l-1} = [e(n-R-O+1), \ldots, e(n-R)]^T$, and $\boldsymbol{e}_l = [e(n-R+1), \ldots e(n)]^T$ is made available for the filter coefficient adaption and for postprocessing, and O being the overlap length.

Being subject to the postfilter, the inherent residual echo $R(l,k)$ and noise signal $N(l,k)$ are suppressed by means of a Wiener postfilter in the frequency domain according to

$$\hat{S}(l,k) = \hat{W}^c_2(l,k) \cdot E^*(l,k), \tag{6.4}$$

with the constrained postfilter coefficients $\hat{W}^c_2(l,k)$ [1]. Based on the unconstrained coefficients $\hat{\boldsymbol{W}}_{2,l} = \left[\hat{\boldsymbol{W}}_2(l,0), \ldots, \hat{\boldsymbol{W}}_2(l,k), \ldots, \hat{\boldsymbol{W}}_2(l,K-1)\right]^T$, a linear constraint is obtained using $\hat{\boldsymbol{w}}_{2,l} = IDFT\left\{\hat{\boldsymbol{W}}_{2,l}\right\}$ to assemble

$$\begin{aligned}\hat{w}^c_{2,l} = {} & [\hat{w}_{2,l}\left(n = K - N_p/2\right), \ldots, \hat{w}_{2,l}(n = K-1), \\ & \hat{w}_{2,l}(n=0), \ldots, \hat{w}_{2,l}\left(n = N_p/2 - 1\right), \boldsymbol{0}^T_{K-N_p}]^T,\end{aligned} \tag{6.5}$$

which contains the linear phase postfilter impulse response of length $N_p \leq K - R - O$. The constrained K-point DFT domain postfilter coefficients are then computed by $\hat{\boldsymbol{W}}^c_{2,l} = DFT\left\{\hat{\boldsymbol{w}}^c_{2,l}\right\}$.

As shown in the coefficient adaption block in Fig. 6.1, the spectral filter coefficients for the echo canceller and Wiener postfilter, $\hat{W}_1(l,k)$ and $\hat{W}_2(l,k)$, are synchronously estimated. This is done by introducing a Markov assumption for the time-varying echo path and exploiting Kalman filter theory [3, 26].

After postfiltering, inverse DFT, and subsequent overlap-add (OLA), the enhanced speech signal $\hat{s}(n)$ is transmitted to the far-end communication partner.

In this setup, an algorithmic delay of $N_p/2 - (R + O) + R = N_p/2 - O$ samples is introduced with $N_p/2$ accounting for the linear phase constrained postfilter, $(R + O)$ being the number of nonzero samples in $IDFT\{\boldsymbol{E}_l\}$ and R being the frame buffer for block processing.

6.3 New Latency-Reduced FDAF with Shadow Filter

6.3.1 Latency Reduction by Postfilter Decimation

Apart from the previously mentioned frame buffering, the linearly constrained postfilter is the only contributor to algorithmic delay. As it can be seen by comparing Figs. 6.1 and 6.2, we now introduce decimation in the DFT domain to reduce the

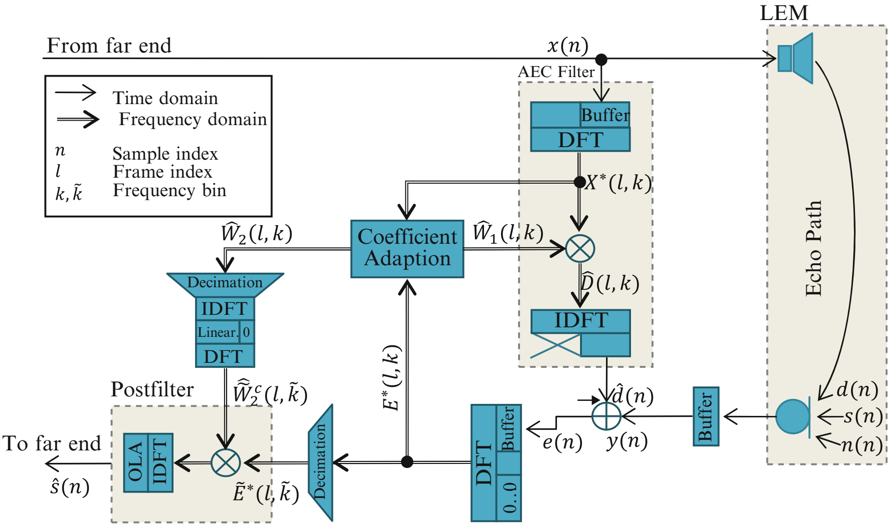

Fig. 6.2 New FDAF-based hands-free system with latency reduction by decimation in the DFT domain

number of DFT bins, which in turn reduces algorithmic delay and computational complexity. In our case, decimation is performed according to

$$\hat{\tilde{W}}_{2,l}\left(\tilde{k}\right) = \begin{cases} \hat{W}_{2,l}(k=0), & \text{for } \tilde{k}=0, \\ \overline{W}_{2,l}\left(\tilde{k}\right), & \text{for } 1 \le \tilde{k} \le \tilde{K}/2-1, \\ \hat{W}_{2,l}(k=K/2), & \text{for } \tilde{k}=\tilde{K}/2, \end{cases} \tag{6.6}$$

with $\overline{W}_{2,l}\left(\tilde{k}\right) = \frac{1}{3}\hat{W}_{2,l}\left(k=2\tilde{k}-1\right) + \frac{1}{3}\hat{W}_{2,l}\left(k=2\tilde{k}\right) + \frac{1}{3}\hat{W}_{2,l}\left(k=2\tilde{k}+1\right)$, and $\hat{\tilde{W}}_{2,l}\left(\tilde{k}\right)$ for $\tilde{k} > \tilde{K}/2$ being defined via the conjugate complex property. Additionally, the decimated DFT error signal $\tilde{E}\left(l,\tilde{k}\right)$ is computed in analogy to $\hat{\tilde{W}}_{2,l}\left(\tilde{k}\right)$. This decimation in the DFT domain therefore leads to a reduced algorithmic delay contribution of the constrained postfilter of $\tilde{N}_p/2-(R+O)$ with $\tilde{N}_p \le \tilde{K}-R-O$ at even lower computational complexity. Furthermore, due to the inherent spectral smoothing of the postfilter coefficients $\hat{\widetilde{\boldsymbol{W}}}^c_{2,l}$, speech quality can be improved by reducing spectral artifacts.

6.3.2 *Shadow-Filter Approach*

The FDAF algorithm shows an excellent double-talk performance even in noisy conditions, without the requirement of an explicit DTD. As a drawback, however, at

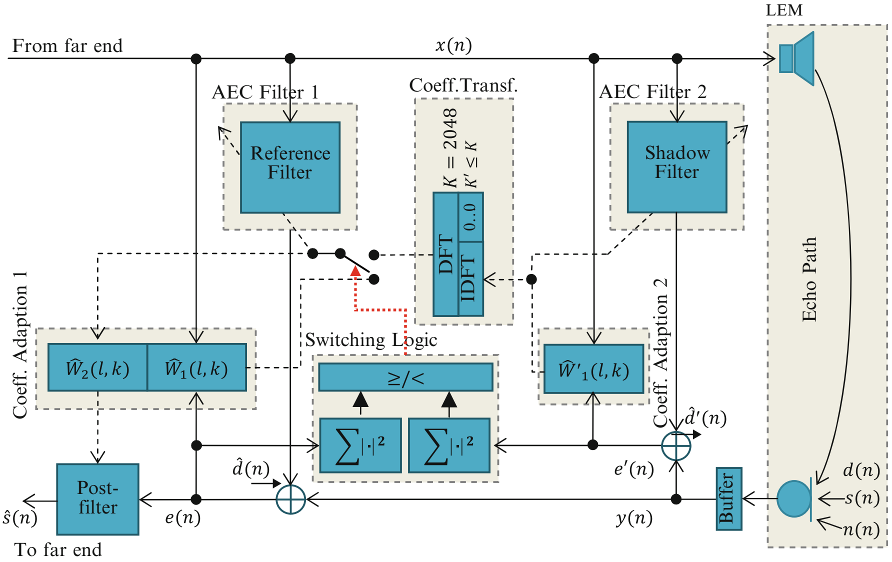

Fig. 6.3 FDAF shadow-filter approach

a sampling rate of 16 kHz, it sometimes cannot achieve a sufficient convergence speed, especially in the case of long filter lengths or highly time-variant echo paths. Luckily, echo path variations in the car cabin are typically not as dynamic as, e.g., in a mobile phone application. Nevertheless, possible solutions for a required faster convergence are the reduction of the filter length (which leads to additional residual echo if the LEM impulse response is not completely covered by the filter) or a faster adaptation rate (which may decrease the double-talk performance). As an alternative, we propose here a shadow-filter approach to overcome this drawback of the FDAF approach [7].

As depicted in Fig. 6.3, we enhance a slowly but accurately converging reference filter (RF) in the foreground with a rapidly converging shadow filter (SF) in the background. The faster convergence of the shadow filter can be achieved by using a shorter filter length and/or different parameters, e.g., for the Markov model of the time-varying echo path [3]. The thereby accelerated shadow filter is better able to follow faster changes of the impulse response and thus leading to a lower error signal energy during highly time-variant time periods. However, care has to be taken to assure robustness against near-end disturbances, since double-talk may be erroneously taken for an IR change.

Changes of the echo path are detected as follows: If the error signal energy of the reference filter $\boldsymbol{e}_l^T \cdot \boldsymbol{e}_l$ is α-times bigger than that of the shadow filter for $1 + L^-$ consecutive frames, a change of the echo path is assumed, and the switching logic, shown in the center of Fig. 6.3, triggers an exchange of filter coefficients in the reference filter (symbolized by the dotted arrow pointing to the switch). In this case, the shadow-filter coefficients $\hat{W}'_1(l,k)$ are expected to better represent the LEM IR. On the other hand, in a time-invariant/slowly changing echo path case, the error signal energy of the reference filter will likely be smaller than that of the shadow filter, and the native reference filter coefficients are used:

$$\hat{\boldsymbol{W}}_{1,l} = \begin{cases} DFT\left\{\left[IDFT'\left\{\hat{\boldsymbol{W}}'^{T}_{1,l}\right\}, \boldsymbol{0}^{T}_{K-K'}\right]\right\}^{T}, & \text{if } \boldsymbol{e}^{T}_{\lambda}\cdot\boldsymbol{e}_{\lambda} > \alpha\, \boldsymbol{e}'^{T}_{\lambda}\cdot\boldsymbol{e}'_{\lambda} \\ & \forall\lambda = l, l-1, \ldots, l-L^{-}; \\ \hat{\boldsymbol{W}}_{1,l}, & \text{else;} \end{cases} \tag{6.7}$$

with $IDFT'\{\cdot\}$ having a reduced length $K' < K$ which stems from the shadow filter.

Please note that the error signal $e'(n)$ of the shadow filter is only deployed to detect changes of the echo path and to adapt the shadow-filter coefficients. In the end, only the error signal of the reference filter $e(n)$, which is either using its native coefficients or a transformed version of the shadow-filter coefficients, is passed over to the postfilter for later transmission.

In so doing, an immediate improvement of the model mismatch can be achieved, leading to a faster convergence. Since the number of DFT coefficients for the reference and shadow filter differs (in our case $K' = K/2$), a transformation of the coefficients has to be performed (shown as "coefficient transformation" block in Fig. 6.3).

6.3.3 Combined Postfilter Decimation and Shadow Filter

In combining both the postfilter decimation of Sect. 6.3.1 and the shadow filter of Sect. 6.3.2 in a new joint approach, advantages of both strategies can be exploited. As proposed before, different parameters were used for the reference filter and the shadow filter to assure good convergence behavior and tracking speed. The whole parameter setting for this joint approach is shown in Table 6.1. This includes the forgetting factor A of the first-order Markov model [3, 6], the AEC filter length $N_w = K - R$ [3, Eq. (20)], the impulse response length of the decimated postfilter that can be chosen to some $\widetilde{N}_p \leq N_p - \widetilde{K}$, and the error power spectral density (PSD) smoothing factor $\lambda_{\phi ee}$. Additionally, a decimation factor of $K/\widetilde{K} = 2$ is chosen. Section 6.4 presents the simulation results of this joint approach.

6.4 Simulations

Our proposed approach has been evaluated by simulation of an LEM system in a car cabin (Volkswagen Touran). Two impulse responses have been measured, originating from both front-door loudspeakers to the car's hands-free microphone

Table 6.1 Parameter settings for the new FDAF approach with decimation, reference filter (RF), and shadow filter (SF)

Description	Value RF	Value SF	Description	Value RF	Value SF
DFT length	$K = 2048$	$K' = 1024$	PF length	$N_p = 1824$	n/a
Dec. DFT length	$\tilde{K} = 1024$	n/a	Dec. PF length	$\tilde{\mathrm{N}}_\mathrm{p} = 800$	n/a
Frameshift	$R = 160$	$R' = 160$	SF loopback	n/a	$L^{-} = 6$
OLA length	$O = 64$	$O' = 64$	SF overestimation	n/a	$\alpha = 3$
Forgetting factor	$A = 0.9995$	$A' = 0.99$	Error PSD smoothing	$\lambda_{\phi ee} = 0.8$	$\lambda'_{\phi ee} = 0.999$
AEC filter length	$N_w = 1888$	$N'_w = 864$			

in the ceiling above the central console. Both measurements were performed in the quiet car with reverberation times of $t_{60} = 35$ ms each. Whereas the front left and the rear right seats were occupied by a quiet passenger during both measurements, the front passenger switched position a bit for the second measurement, but keeping a typical driving position in both cases.

These two impulse responses were used to compute the echo signal $d(n)$ of the far-end speaker, shown as waveform in the upper part of Figs. 6.4 and 6.5, by convolution with the far-end signal $x(n)$. The used impulse response was switched from the first measurement to the second after 25 s.

The near-end speech signal $s(n)$ waveform in double-talk is shown in front of the echo signal waveform in Fig. 6.5. Speech signals are concatenated samples of the NTT wideband speech database. Whereas the male far-end speaker is continuously active during both the single- and double-talk scenarios (Figs. 6.4 and 6.5), the female near-end speaker is only intermittently active during the double-talk scenario (Sect. 6.5). All simulations are performed at an input signal-to-echo ratio of SER = 0 dB.

For evaluation of the performance of the underlying hands-free system, different instrumental measures have been used. For assessment of the system's ability to suppress the echo signal, echo return loss enhancement (*ERLE*) plots for the single- and double-talk scenario are given in Figs. 6.4 and 6.5 as lines above the waveforms. *ERLE* is defined and recursively estimated as follows:

$$\begin{aligned} ERLE(n) &= \frac{E\{d^2(n)\}}{E\Big\{\big(d(n) - \hat{d}(n)\big)^2\Big\}} \\ &\approx \frac{(1-\beta)d^2(n) + \beta d^2(n-1)}{(1-\beta)\big(d(n) - \hat{d}(n)\big)^2 + \beta\big(d(n-1) - \hat{d}(n-1)\big)^2} \end{aligned}$$

with smoothing factor $\beta = 0.9996$.

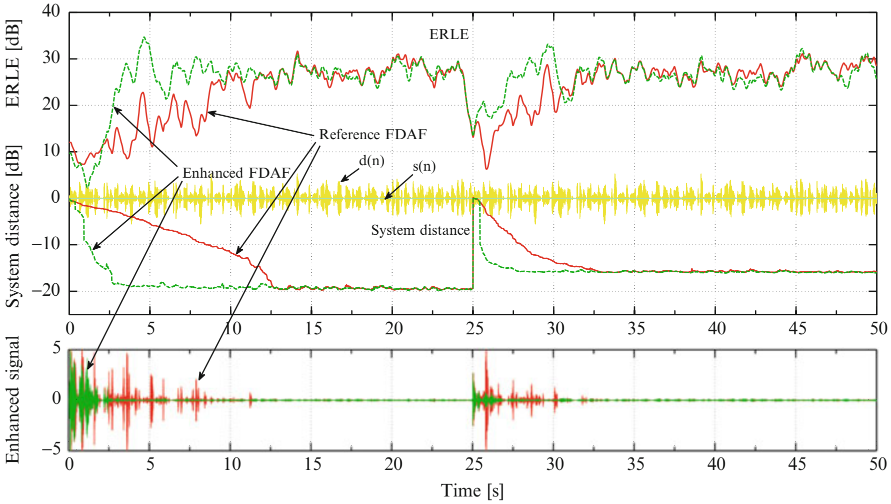

Fig. 6.4 Reference FDAF algorithm (*solid lines*) and enhanced FDAF with postfilter decimation and shadow-filter approach (*dashed lines*) in *single-talk*. The IR changes at 25 s. *Above*: ERLE (*top*) and system distance (*bottom*) for single-talk signal $d(n)$. *Below*: Enhanced microphone signals

To evaluate the convergence and tracking performance in single- and double-talk, also the normalized system distance is shown in Figs. 6.4 and 6.5, defined as

$$\frac{||h_{\Delta}{}^2||}{||h_i{}^2||} = \frac{||h_i - \hat{w}_{1,l}{}^2||}{||h_i{}^2||} \tag{6.8}$$

with h_i, $i = 1, 2$, being one of the two measured impulse responses, and $\hat{w}_{1,l} = IDFT\{\hat{W}_{1,l}\}$.

For both the *ERLE* and system distance plots, solid lines correspond to the reference setup and dashed lines correspond to the enhanced system, making use of shadow filtering and decimation. In the lower part of Figs. 6.4 and 6.5, the enhanced signals $\hat{s}(n)$ are shown.

For evaluation of the degradation of the wideband (uplink) speech component, the perceptual evaluation of speech (PESQ) measure according to ITU-T recommendation P.862.2 [30] is used, yielding objective listening-quality mean opinion scores (MOS_{LQO}). The so-called MOS_{LQO}^{PF} score is used here to evaluate only the quality degradation of the speech *component* by *postfiltering*, marked by the superscript “PF”.

6.4.1 Far-End Single-Talk Scenario

As it can be easily seen in the upper part of Fig. 6.4, the continuous far-end speech input leads to a final normalized system distance of −19 dB and −16 dB for the two

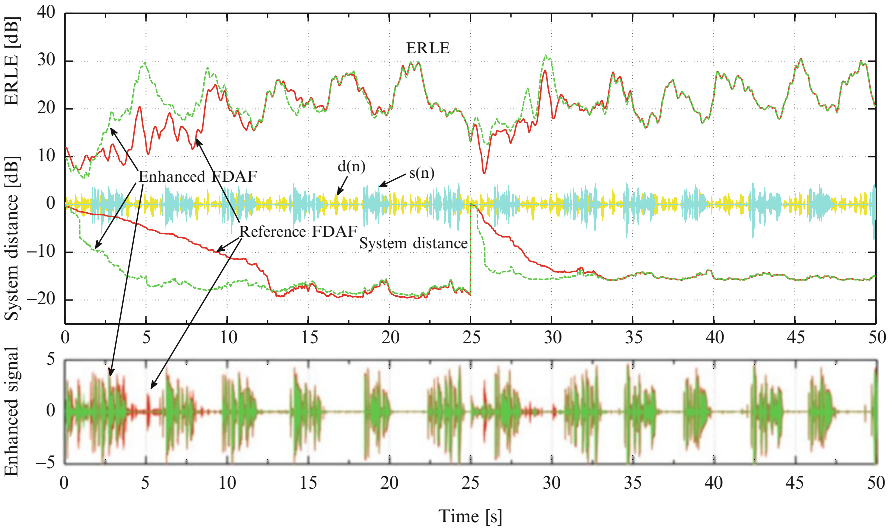

Fig. 6.5 Reference FDAF algorithm (*solid lines*) and enhanced FDAF with postfilter decimation and shadow-filter approach (*dashed lines*) in *double-talk*. The IR changes at 25 s. *Above*: ERLE (*top*) and system distance (*bottom*) for double-talk signal $d(n) + s(n)$. *Below*: Enhanced microphone signals

measured IRs, respectively. Due to the quite long adaptive filter length of $N_w = K - R = 2048 - 160 = 1888$ taps, adaptation of the reference filter (shown as solid lines in Fig. 6.4) is quite slow. A system distance of −10 dB is reached 8.5 s after initialization and 3 s after the IR switch. Since this convergence and tracking time is simply not sufficient, shadow filtering has been used to achieve faster filter adaptation. In addition, decimation is used for the postfilter. The *ERLE* and system distance plots of this enhanced setup are shown as dashed lines in the upper part of Fig. 6.4. These measures lead to significantly reduced convergence times after initialization and for tracking of abrupt IR changes. In both cases, convergence time is reduced to less than 1 s to reach −10 dB system distance.

Table 6.2 shows the mean *ERLE*, mean system distance, and algorithmic delay for four different approaches. The mean values have been computed for the signals as they are shown in Fig. 6.4. However, simulations based on different NTT datasets show comparative results. The reference algorithm, as described in Sect. 6.2, is shown in the first row. Due to the quite long convergence time of this approach, the mean *ERLE* and system distance values in the evaluated period of time significantly differ from the mean values of the fully converged filter. The algorithmic delay of $(N_p/2 - (R + O) + R) \cdot 1/16\,\text{kHz} = 53$ ms of this approach is mainly accounted to the delay introduced by the Wiener postfilter (6.4) which amounts to $(N_p/2 - (R + O)) \cdot 1/16$ kHz = 43 ms. The remaining $R \cdot 1/16$ kHz = 10 ms delay is caused by the buffering, which is necessary for the block processing in the frequency domain (see Fig. 6.1).

Table 6.2 Performance evaluation in single-talk

	$\overline{\text{ERLE}}$	$\overline{\text{SYSDIS}}$	Delay	$\text{MOS}_{\text{LQO}}^{\text{PF}}$
Reference	23.6 dB	−13.7 dB	53 ms	n/a
Ref. + SF.	25.5 dB	−16.6 dB	53 ms	n/a
Ref. + Dec.	23.7 dB	−13.7 dB	21 ms	n/a
Ref. + Dec. + SF.	25.7 dB	−16.8 dB	21 ms	n/a

The introduction of shadow filtering, as it is depicted in Fig. 6.3, leads to better convergence behavior. This can also be seen by looking at the improved mean values for *ERLE* and system distance in the second row of Table 6.2. In this special case the mean *ERLE* could be increased by around 2 dB, and the normalized system distance could be decreased by nearly −3 dB. The algorithmic delay remains unchanged.

Further improvements can be achieved by decimation of the DFT coefficients for the spectral gain and input signal of the postfilter, as it is shown in Fig. 6.2. In so doing, three improvements can be achieved simultaneously: First, due to the halved DFT and IDFT lengths in the postfilter, the computational complexity is somewhat reduced; the decimation in the frequency domain acts as smoothing of the postfilter weights and the input speech vector; this is leading to a better residual echo suppression; and as third factor, quality degradation of the speech component can be reduced, as will be shown in Sect. 6.4.2. However, decimation should only be introduced carefully. The effect of this decimation together with a shadow filter can be seen by looking at the results in the fourth row of Table 6.2. *ERLE*, and system distance can be further improved by 0.2 dB and −0.2 dB, compared to the shadow-filter-only approach, shown in the second row of Table 6.2. Whereas these improvements are rather small, the important effect of a much smaller algorithmic delay of 21 ms compared to the former 53 ms is achieved by the considerably smaller length of the postfilter. By only applying postfilter decimation, as shown in the third row, algorithmic delay remains low at 21 ms, with *ERLE* and system distance values being comparable to the reference.

6.4.2 Double-Talk Scenario

Some effects become even more clear when regarding a double-talk scenario, as it is shown in Fig. 6.5. The presence of near-end speech or noise is posing an interference to the adaptive filter, hence leading to slower convergence or misadaptation to the interfering signal. However, looking at Fig. 6.5, the double-talk performance of the FDAF algorithm can still be considered as excellent. The *ERLE* values drop by around −8 dB during double-talk but still keep a minimum value of around 20 dB in the converged state. Furthermore, due to this high robustness, convergence times more or less stay the same as it can be seen by looking at the system distance plot of the reference filter in Fig. 6.5.

Table 6.3 Performance evaluation in double-talk

	$\overline{\text{ERLE}}$	$\overline{\text{SYSDIS}}$	Delay	$\text{MOS}_{\text{LQO}}^{\text{PF}}$
Reference	20.1 dB	−13.0 dB	53 ms	3.3
Ref. + SF.	20.9 dB	−15.0 dB	53 ms	3.3
Ref. + Dec.	20.1 dB	−13.0 dB	21 ms	3.6
Ref. + Dec. + SF.	21.3 dB	−15.3 dB	21 ms	3.6

Looking at the upper part of Fig. 6.5 (*ERLE*), it can be seen that the shadow-filter-enhanced approach is still performing well in a double-talk scenario. In this case, convergence time to reach a system distance of −10 dB could be reduced to about 2 s. For both scenarios, this convergence time is of course dependent on the speech signals as well as on proper tuning of the shadow-filter parameters.

Comparing the mean *ERLE* and system distance values from the single-talk scenario (Table 6.2) to the double-talk scenario (Table 6.3), an expected, yet moderate, performance degradation is observed. For the reference approach, shown in the first row, the mean *ERLE* value drops to 20.1 dB, and the system distance slightly increases to −13.0 dB. Of course, algorithmic delay stays at 53 ms. As introduced before, perceptual quality degradation by the postfilter is evaluated by the MOS_{LQO}^{PF}. In this case, a fair to good score of 3.3 is achieved.

By making use of the shadow filter, both the mean *ERLE* and mean system distance values could be improved. $\overline{ERLE}$ increases to 20.9 dB, whereas $\overline{SYSDIS}$ improves to −15.0 dB. Algorithmic delay as well as MOS_{LQO}^{PF} remains constant when shadow filtering is applied.

Additional application of the decimation approach is again able to further improve the performance, as the results in the fourth row show. Here, $\overline{ERLE}$ increases to 21.3 dB, whereas $\overline{SYSDIS}$ slightly improves to −15.3 dB. This approach additionally offers two further advantages: As already shown in Table 6.2, algorithmic delay is reduced to 21 ms, and it also leads to an improvement of the perceptual quality of the speech component. This can be seen from the surprising fact that the MOS_{LQO}^{PF} score improves from 3.3 (without decimation) to 3.6 (with decimation). Again, postfilter decimation alone, as shown in the third row, does not show different *ERLE* or system distance results compared to the reference, whereas algorithmic delay and the MOS_{LQO}^{PF} score are improved.

6.5 Conclusions

We have presented a wideband automotive hands-free system for mobile HD Voice services. It is based on a shadow-filter approach for an FDAF-based acoustic echo canceller, which is significantly improving the convergence speed and tracking performance. Our approach excels in double-talk performance, revealing a high quality of the speech component in uplink direction. By decimation of the frequency domain postfilter coefficients, the computational complexity, algorithmic delay, and perceptual speech quality could be improved. Experimental results show a good performance in a simulated car environment.

References

1. J.J. Shynk, Frequency-domain and multirate adaptive filtering. IEEE Signal Process. Mag. **9**(1), 14–37 (1992)
2. C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, J. Tilp, Acoustic echo control. An application of very-high-order adaptive filters. IEEE Signal Process. Mag. **16**(4), 42–69 (1999)
3. G.W. Enzner, P. Vary, Frequency-domain adaptive Kalman Filter for acoustic echo control in hands-free telephones. Signal Process. Elsevier **86**, 1140–1156 (2006)
4. H.W. Löllmann, P. Vary, Uniform and low delay filter-banks for speech enhancement. EURASIP Speech Commun. **49**(7–8), 574–587 (2007)
5. H. Puder, P. Dreiseitel, Implementation of a hands-free car phone with echo cancellation and noise-dependent loss control, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 6 (2000), pp. 3622–3625
6. S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 4 edn., (2002)
7. E. Hänsler, G. Schmidt, *Acoustic echo and noise control: a practical approach* (Wiley, Hoboken, NJ, 2004)
8. N.K. Jablon, On the complexity of frequency-domain adaptive filtering. IEEE Trans. Signal Process. **39**(10), 2331–2334 (1991)
9. C. Beaugeant, M. Schönle, I. Varga, Challenges of 16 kHz in acoustic pre- and post-processing for terminals. IEEE Commun. Mag. **44**(5), 98–104 (2006)
10. B. Widrow, P. N. Stearns, *Adaptive Signal Processing*. Prentice Hall, 1 edn.,(1985)
11. A. H. Sayed, *Fundamentals of Adaptive Filtering*. Wiley-IEEE Press, 1 edn., (2003)
12. S. Gustafsson, R. Martin, P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony. Signal Process. **64**(1), 21–32 (1998)
13. C. Yemdji, M. Idrissa, N. Evans, C. Beaugeant, Efficient low delay filtering for residual echo suppression, in *European Signal Processing Conference* (*EUSIPCO*), (2010)
14. Y. Ephraim, D. Malah, Speech enhancement using optimal non-linear spectral amplitude estimation, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 8 (1983), pp. 1118–1121
15. T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Appl. Signal Process. **7**, 1110–1126 (2005)
16. P. Scalart, J. Filho, Speech enhancement based on a priori signal to noise estimation, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 7 (1996), pp. 629–632
17. H.-C. Shin, A. Sayed, Variable step-size NLMS and affine projection algorithms. IEEE Signal Process. Lett. **11**, 132–135 (2004)
18. K. Ozeki, T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties. Electron. Commun. Jpn. **67**(5), 19–27 (1984)
19. S. Gay, S. Travathia, The fast affine projection algorithm, in *Proceedings IEEE International Conference Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 3 (1995), pp. 3023–3027
20. J. Cioffi, T. Kailath, Fast, recursive-least-squares transversal filters for adaptive filtering. IEEE Trans. Acoustics Speech Signal Process. **32**(2), 304–337 (1984)
21. R.E. Kalman, A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**, 35–45 (1960)

22. K. Steinert, M. Schönle, C. Beaugeant, T. Fingscheidt, Hands-free system with low-delay subband acoustic echo control and noise reduction, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing* (*ICASSP*), (2008) pp. 1521–1524
23. R. Crochiere, L.R. Rabiner, *Multirate digital signal processing* (Prentice Hall, Englewood Cliffs, NJ, 1983)
24. J. Shynk, R. Gooch, Frequency-domain adaptive pole-zero filtering. Proc. IEEE **73**, 1526–1528 (1985)
25. G. W. Enzner, *A model-based optimum filtering approach to acoustic echo control*: *theory and practice*. PhD thesis, Rheinisch-Westfälische Technische HochschuleAachen, 2006
26. G. W. Enzner, P. Vary, Robust and elegant, purely statistical adaptation of acoustic echo canceler and postfilter, in *International Workshop on Acoustic Echo and Noise Control* (*IWAENC*), (Kyoto, 2003), pp. 43–46
27. M. Ihle, K. Kroschel, Integration of noise reduction and echo attenuation for handset-free communication, in *International Workshop on Acoustic Echo and Noise Control* (*IWAENC*), (1997) pp. 69–72
28. E. Hänsler, U. Schmidt, Hands-free telephones—joint control of echo cancellation and postfiltering. IEEE Signal Process. Mag. **80**, 2295–2305 (2000)
29. C. Beaugeant, V. Turbin, P. Scalart, A. Gilloire, New optimal filtering approaches for hands-free telecommunication terminals. Signal Process. **64**(1), 33–47 (1998)
30. ITU-T, Rec. P.862.2: *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, (2007)