Low delay filtering for joint noise reduction and residual echo suppression

Christelle Yemdji, Moctar I. Mossi and Nicholas Evans
EURECOM Institute
06560 Sophia-Antipolis, France
{yemdji, mossi, evans}@eurecom.fr
web: www.eurecom.fr

Christophe Beaugeant
Infineon Technologies
06560 Sophia-Antipolis, France
christophe.beaugeant@infineon.com
web: www.infineon.com

***Abstract*—In mobile terminals speech quality is often degraded by acoustic echo and noise. Existing approaches to tackle these problems involves separated solutions. However, combined solution have been proposed recently. Low signal delay and reduced computational load are the main arguments in favor of joint noise and echo processing. Most such algorithms operate in the frequency domain. In recent works, the performance of low delay filtering structures, which could be used as alternatives to spectral weighting, have been studied for noise reduction and echo suppression separately. In this paper, we investigate these low delay approaches for joint noise reduction and echo suppression. Results show that the approach based on the inverse discrete Fourier transform performs as well as alternative approaches but with lower signal delay and reduced computational complexity.**

***Index Terms* — Echo suppression, noise reduction, sub-band filtering, FIR filter.**

I. Introduction

An acceptable level of speech quality is an important requirement for any telecommunications terminal. With mobile devices, however, speech quality is often degraded by varying levels of ambient noise and acoustic echo. Acoustic echo results from the coupling between the loudspeaker and the microphone. The far-end speaker sometimes hears a delayed version of their own voice, where the delay is introduced by the communications link. In noisy environments, the microphone is sensitive to near-end speech and ambient noise which are both transmitted to the far-end speaker. Acoustic echo cancellation (AEC) and noise reduction (NR) are used to tackle these problems [1].

Most approaches to AEC are based on adaptive filters [1]. As illustrated in Figure 1, an adaptive filter is used to generate an estimate of the echo signal which is then subtracted from the microphone signal. However, because of the limited filter order, changes in the acoustic path and non-linearities, some residual echo often remains. Postfilters are commonly used to obtain further echo attenuation [1]. Sub-band echo postfilters have been widely investigated and have proven to be a good compromise between efficient echo suppression and low computational complexity [1], [2], [3].

Noise reduction algorithms usually operate in the frequency domain and are generally based on the assumption that noise is an additive and relatively stationary perturbation. Commonly used noise reduction algorithms are based on an estimate of the noise signal which is used to calculate a noise reduction filter [1], [4], [5].

Residual echo suppression and noise attenuation problems can be considered independently. Attempts to build combined systems for noise reduction and echo suppression can be found in the literature [2], [6] and have shown to perform well. The main arguments in favor of such a combined system are the reduced computational complexity and lower signal delay due to the use of one analysis and synthesis filter bank, instead of two, as the filtering takes place in the frequency domain. Further delay reductions can be achieved by filtering degraded speech signals in the time domain with finite impulse response (FIR) filters. Approaches to calculate such an FIR filter include the Filter Bank Equalizer (FBE), the Low Delay Filter (LDF), both presented in [7], and the inverse discrete Fourier transform (IDFT) of spectral gains [1], [8]. In [7], the performances of the FBE and LDF approaches were assessed for noise reduction. In [9], we investigated the performances of FBE, LDF and IDFT filters for sub-band echo postfiltering.

In this paper, we compare these low delay filtering structures with the classic spectral weighting for combined noise reduction and echo suppression. We show that FIR filters can be efficiently used for noise reduction and echo suppression. An emphasis is made on the IDFT approach which is the least computational demanding and still yields good performance.

This paper is organized as follows. In the next section we present the algorithm used for the calculation of the spectral gains and different filtering schemes which are compared. In Section III we describe our experimental setup and present our findings. Our conclusions are presented in Section IV.

II. System description

Figure 1 shows the speech enhancement scheme used in our investigations: AEC followed by noise reduction and residual echo suppression, here combined as a postfilter. The microphone signal $y(n)$ is composed of the near-end speech signal $s(n)$, the echo signal $d(n)$ and the noise signal $n(n)$. The adaptive filter is used to generate an estimate of the echo signal $\hat{d}(n)$ which is subtracted from the microphone signal to obtain the error signal $e(n)$. This error signal is composed of the residual echo $d_r(n)$, the near-end speech $s(n)$ and the noise signal $n(n)$. The postfilter then aims to suppress the residual echo and to attenuate noise.

In the following, we describe the joint noise reduction and

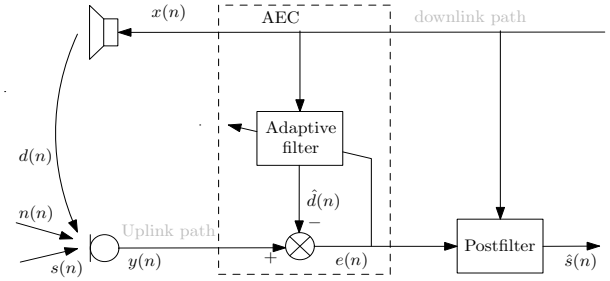
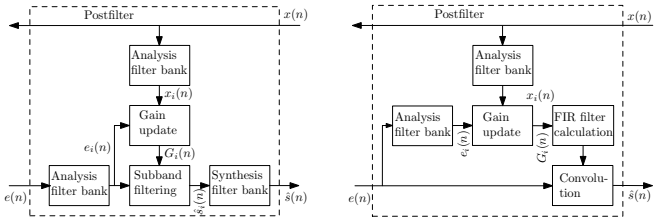


Fig. 1. Speech enhancement scheme illustrating AEC followed by a postfilter. Here, the postfilter is a joint noise reduction and residual echo suppression module.



(a) Perturbations filtering in the frequency domain.

(b) Perturbations filtering with an FIR filter.

Fig. 2. Postfiltering scheme.

echo postfilter that are investigated in this paper. Section II-A details the sub-band analysis used. In Section II-B, we present the algorithm used to calculate the postfilter spectral gains. Lastly, Section II-C presents the 4 different filtering approaches compared.

A. Sub-band analysis

As shown in Figure 2, the error signal $e(n)$ and the loudspeaker signal $x(n)$ are split into sub-band signals $e_i(n)$ and $x_i(n)$ respectively, where i denotes the sub-band index and ranges from 0 to $M-1$. In our system, sub-band analysis and synthesis are performed through a discrete Fourier transform-modulated filter bank. One property of such filter banks is that each bandpass filter corresponds to a frequency shifted duplicate of a lowpass filter $h(n)$. In the literature $h(n)$ is referred to as a prototype filter [1].

The number of sub-bands M is set to 64 and a dowsampling factor r is set to 32. The length L of the prototype filter is set to 128. A similar sub-band analysis was reported in [9].

B. Noise reduction and residual echo suppression

The postfilter spectral gains used to process degraded speech signals are defined as the product of the noise reduction and echo suppression gains:

$$G_i(n) = G_i^n(n).G_i^{d_r}(n), \tag{1}$$

where i is the sub-band number, $G_i^n(n)$ is the noise reduction gain and $G_i^{d_r}(n)$ is the residual echo suppression gain. The noise reduction and echo postfiltering spectral gains are calculated independently. The echo postfilter is updated using a Wiener rule for echo suppression whereas the noise reduction filter is a low complexity noise reduction algorithm [10]. The noise reduction rule is based on the assumption that the amount of noise that should be attenuated is proportional to the signal to noise ratio (SNR).

1) Echo postfiltering rule: The gain of the echo postfilter is the same as that used in [9] and is updated as follows [2]:

$$G_i^{d_r}(n) = \frac{\xi_i(n)}{1+\xi_i(n)}, \tag{2}$$

where $\xi_i(n)$ is the signal (near-end speech) to echo ratio (SER). In our implementation, the SER is estimated through the Ephraim and Malah approach [11]:

$$\xi_i(n) = \beta.\frac{\hat{s}_i^2(n-1)}{\hat{\gamma}_i^{d_r d_r}(n-1)} + (1-\beta).max(\xi_i^{post}(n), 0) \tag{3}$$

where the smoothing constant β lies in the interval $]0,1[$, $\hat{s}_i(n-1)$ is the i^{th} sub-band near-end speech signal estimate, $\hat{\gamma}_i^{d_r d_r}(n)$ is the residual echo spectral density and $\xi_i^{post}(n)$ is the a posteriori SER. The residual echo spectral density in Equation 3, $\hat{\gamma}_i^{d_r d_r}(n)$, is estimated according to [2]:

$$\hat{\gamma}_i^{d_r d_r}(n) = \frac{\gamma_i^{xe}(n)}{\gamma_i^{xx}(n)}, \tag{4}$$

where $\gamma_i^{xe}(n)$ is the crosspower spectral density between $x(n)$ and $e(n)$ and $\gamma_i^{xx}(n)$ is the loudspeaker power spectral density. The a posteriori SER in Equation 3, $\xi_i^{post}(n)$, is calculated according to:

$$\xi_i^{post}(n) = \frac{e_i^2(n)}{\hat{\gamma}_i^{d_r d_r}(n)} - 1. \tag{5}$$

The spectral densities $\gamma_i^{xx}(n)$ and $\gamma_i^{xe}(n)$ are estimated through autoregressive smoothing as in [2]. To avoid artifacts, the echo reduction gain is limited to a spectral floor which is adapted proportionally to the noise level.

2) Noise reduction rule: The noise reduction gains used are updated as follows:

$$G_i^n(n) = \min(\alpha.\chi_i^{\lambda}(n), 1), \tag{6}$$

where α and λ are empiricallly optimised constants and $\chi_i(n)$ is the signal (near-end speech) to noise ratio (SNR). The SNR is estimated as follows:

$$\chi_i(n) = \frac{\gamma_i^{ee}(n)}{\hat{\gamma}_i^{nn}(n)}, \tag{7}$$

where $\hat{\gamma}_i^{nn}(n)$ is the estimate of the noise spectral density, which is obtained by minimum statistics tracking method [5]. The spectral density $\gamma_i^{ee}(n)$ is estimated through autoregressive smoothing as in [2]. To avoid artifacts, the noise reduction gain is limited to a fixed spectral floor.

C. Filtering approaches

As mentioned earlier, the filtering of the degraded speech signals can take place in the sub-band domain or, alternatively, in the time domain.

Filtering in the frequency domain consists in applying the postfilter spectral gains to the sub-band signals through a

multiplication [1] (see Figure 2(a)). The fullband microphone signal $\hat{s}(n)$ is then recovered by processing the sub-band signals $\hat{s}_i(n)$ through a synthesis filter bank. In the remainder of this paper, this approach will be referred to as the SF (sub-band filtering) approach. For filtering in time domain, the postfilter sub-band gains are transformed into an FIR filter before being applied to the fullband microphone signal as shown in Figure 2(b). Methods to determine the FIR filter include the FBE approach, the LDF approach [7] and the IDFT of the spectral gains [1], [9]. The FBE approach is the time domain mathematical equivalent of sub-band filtering. In this case, the length of the FIR filter is equal to that of the prototype filter, which here is equal to 128. The LDF approach is derived from the FBE by truncating its impulse response with a shorter window. In our implementation, its length is truncated to 64. The IDFT filter is an intuitive approach which consists in using the IDFT of the postfilter spectral gains to obtain a time domain filter [1]. The length of the IDFT filter is equal to the number of sub-bands which is 64 in our case.

III. Experimental work

A. Experimental setup

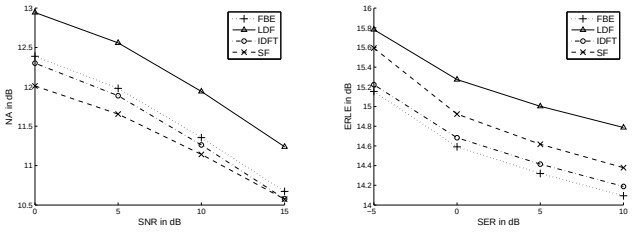
The postfilter described in Section II is assessed through simulations with speech signals. As shown in Figure 1, the microphone signal is first processed by an AEC module. The AEC algorithm used in our simulations is a sub-band normalized least mean square algorithm [9] as in our earlier work in [9].

Our test database is generated using a set of four far-end speech signals and four near-end speech signals. The microphone signals used in our simulations contain near-end speech only, echo only and double talk periods with either car, cafe or babble noise. The echo signal is obtained by convolving the loudspeaker signals with an acoustic path measured from real mobile terminals in an office environment. The loudspeaker and near-end speech levels are both set to -26dB using the ITU-T implementation of the speech voltmeter [12] and the different echo and noise levels are also set using the same tool. The SNR ranges from 0 to 15dB while the SER ranges from -5 to 10dB. Our database of degraded speech signals contains 192 sets of microphone and loudspeaker signals.

Performance of the different filtering approaches is assessed through objective measurements and informal listening tests. Echo suppression is assessed in terms of echo return loss enhancement (ERLE). Noise reduction is assessed in terms of noise attenuation (NA). Perturbation (noise or echo) attenuation is measured over adjacent windows of N samples:

$$D(m) = 10.log_{10}\Big(\frac{\sum_N e(n)^2(N)}{\sum_N \hat{s}^2(N)}\Big) \qquad (8)$$

where $D(m)$ stands for ERLE or NA and N spans over 256 samples. ERLE and NA are computed according to Equation 8. However, ERLE is measured during echo only periods while NA is measured during noise only periods. Speech distortion is assessed in terms of the cepstral distance (CD) measurement



(a) NA against SNR (b) ERLE against SER

Fig. 3. Perturbation attenuation

between the clean speech $s(n)$ and the weighted speech signal $\bar{s}(n)$ [13] as follows:

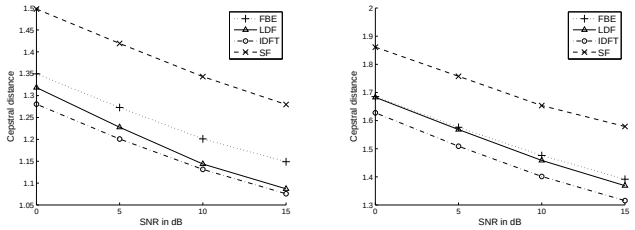
$$\begin{aligned} C_s(m) &= IDFT\{ln|DFT(s(N))|\} \\ CD(m) &= \sqrt{\sum_N [C_s(m) - C_{\bar{s}}(m)]^2}. \end{aligned} \qquad (9)$$

The weighted speech signals $\bar{s}(n)$ are obtained with a method similar to [13]. When processing degraded speech signals, the updated spectral gains $G_i(n)$ are stored. These gains are applied to the clean near-end speech $s(n)$ to obtain the weighted speech signal $\bar{s}(n)$.

B. Results

Figure 3(a) shows NA against SNR. The NA curves show that the LDF approach achieves the best performance in terms of noise reduction. The SF method achieves the worst performance but there is little difference between the SF, FBE and IDFT approaches. In general, all the different filtering approaches have very similar results: the differences between NA curves is less than 1dB. The ranking of the FBE, LDF and SF performance is the same as in [7]. The novelty, over what is already presented in [7], is the IDFT approach which, as we see in Figure 3(a), is as effective as the other approaches in reducing noise.

Figure 3(b) shows ERLE against SER. All the filtering approaches have very similar results (differences between ERLE curves is less than 0.6dB). Here, we can highlight two main differences from results presented in our previous work [9]. First, in this paper the FBE achieves the worst performance and second, the performance for the IDFT approach curve is very close to that of the SF approach. Tests on clean speech signals (no additive noise) using the system described in this paper confirm the observations in [9] concerning the performance of the IDFT approach. The observations in [9] were that the IDFT filter had the worst performance. We explained this poor performance by the fact that the effective frequency response of the IDFT filter had, in that case, large variations (Gibbs phenomenon) between consecutive sub-bands where the difference in gain was high (greater than 10dB). Our system is designed so that the difference in gain between consecutive sub-bands cannot be higher than the spectral floor which depends on the noise level (see Sections II-B1 and II-B2). In [9], the difference in gain between consecutive sub-bands can sometimes be large because there is no noise. The work presented here tackles the problem of echo in the



(a) CD during near-end speech (b) CD during double talk

Fig. 4. Cepstral distance

presence of noise, and this leads to a reduction in the difference in gain between consecutive sub-bands and thus to better results for the IDFT approach.

Figure 4(a) shows cepstral distance against SNR during near-end speech only periods, i.e. distortions resulting from noise reduction. We see that the SF approach brings the most distortions whereas the IDFT approach brings the least. These results are different from those presented in [7] in which the SF and FBE approaches were reported to produce speech of equivalent quality. Our explanation is that this difference is due to the analysis and synthesis filterbanks which are not the same in both papers. In [7], the analysis and synthesis filterbanks used for the SF and FBE approaches are not the same whereas in this paper they are. Figure 4(b) shows CD against SNR during double talk periods. The ranking of the different filtering methods remains the same as in Figure 4(a). We note an increase of the CD values because during double talk both echo and noise reduction are active. We also see that the CD values for the FIR filters are very close to each other.

Informal listening tests reveal that near-end speech during double talk periods is distorted whereas no distortion is noticed during near-end speech periods. Listening tests with weighted speech signals $\bar{s}(n)$ reveal the presence of small distortions of near-end speech during near-end speech only periods. These observations imply that echo processing brings more distortions than noise reduction no matter the filtering approach used. Distortions introduced by the noise reduction are not audible in processed speech signals due to the masking effect of residual noise present in processed speech signals. The distortions observed are crackling noises for signals processed by FIR filters and the presence of musical noise (random spectral peaks of short duration) for signals processed in the spectral domain. As explained in [9], the crackling comes from the fact that the frequency response of FIR filters is smoother than that of the original spectral gains which are defined per sub-band. Nevertheless, these distortions were not perceived as annoying even during double talks periods as they are masked by residual noise. The differences between signals processed by the IDFT, LDF and FBE approaches were hardly audible. This confirms what might be expected on account of results illustrated in Figure 4.

IV. Conclusion

This paper presents the first comparison of four different filtering approaches that can be used for joint echo and noise reduction. Performance of the different filtering approaches is compared through objective measurements and informal listening tests.

In the system described here, the FBE has a lower delay and an increased computational complexity than the SF. In our system, the LDF and IDFT approaches are equivalent in terms of signal delay and speech quality. However, the IDFT approach has the advantage of low complexity and is thus an appealing alternative to the SF approach.

Results showed that all the filtering methods studied here can be used efficiently for noise reduction and echo suppression with few differences between signals processed in the sub-band domain and those processed with FIR filters. Substituting the classic SF approach by one of the FIR filters presented here has an impact on speech quality as we trade-off one artifact against another. As has already been done for musical noise reduction, it is of interest to study means of reducing the crackling noise artifacts in order to improve speech quality of signals processed by FIR filters.

References

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

[2] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Processing*, vol. 64, no. 1, pp. 33–47, 1998.

[3] E. Habets, S. Gannot, I. Cohen, and P. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1433 – 1451, November 2008.

[4] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 799 –807, November 2001.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504 – 512, July 2001.

[6] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Hands-free system with low-delay subband acoustic echo control and noise reduction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2008, pp. 1521–1524.

[7] H. W. Löllmann and P. Vary, "Uniform and warped low delay filterbanks for speech enhancement," *Speech Communications*, vol. 49, no. 7–8, pp. 574–587, 2007.

[8] E. Hänsler and G. U. Schmidt, "Hands-free telephones - joint control of echo cancellation and postfiltering," *Signal Processing*, vol. 80, no. 11, pp. 2295–2305, 2000.

[9] C. Yemdji, M. Mossi Idrissa, N. W. D. Evans, and C. Beaugeant, "Efficient low delay filtering for residual echo suppression," in *2010 European Signal Processing Conference (EUSIPCO-2010)*, Aalborg, Denmark, 8 2010.

[10] P. Degry and C. Beaugeant, "Solution to speech quality improvement in telecommunication terminals," in *ITG Fachtagung Sprachkommunikation*, October 2008.

[11] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1983, pp. 1118–1121.

[12] ITU-T, "ITU-T recommendation P.56: objective measurement of active speech level," 1993.

[13] T. Fingscheidt and S. Suhahi, "Quality assessment of speech enhancement systems by separation of enhanced speech, noise and echo," in *Proc. Interspeech*, 2007, pp. 818 – 821.