DUAL CHANNEL ECHO POSTFILTERING FOR HANDS-FREE MOBILE TERMINALS

Christelle Yemdji[1]*, Moctar Mossi Idrissa*[1]*, Nicholas Evans*[1]*, Christophe Beaugeant*[2] *and Peter Vary*[3]

[1]EURECOM, Multimedia Department, Sophia-Antipolis, France
[2]Intel, Mobile Comunications Group, Sophia-Antipolis, France
[3]Institute of Communication Systems and Data Processing (IND), RWTH Aachen, Germany
{yemdji, mossi, evans}@eurecom.fr christophe.beaugeant@intel.com vary@ind.rwth-aachen.de

ABSTRACT

In mobile communications, speech is often degraded by ambient noise and acoustic echo. Both problems have attracted a high level of research interest over the last decades. Recently dual channel solutions for noise reduction have been investigated and can yield better performance than single channel solutions. This paper presents an analysis of the echo problem based on recordings with a hands-free dual channel mock-up phone and proposes a novel residual echo power spectral density (PSD) estimator that uses two microphone signals instead of one. The proposed PSD estimate is compared to an existing single channel residual echo PSD estimator before being assessed within an echo postfilter. Our experiments show that the proposed dual-channel echo PSD outperforms the single channel postfilter, especially at low signal to echo ratios.

Index Terms— echo postfiltering, dual channel, relative transfer function

1. INTRODUCTION

With increased flexibility and mobility, mobile terminals are one of the most popular and widespread telecommunications terminals. Mobile terminals are very likely to be used in very different and adverse conditions such as in hands-free mode or in noisy environments. In hands-free mode, part of the far-end voice signal played by the loudspeaker is picked by the microphone. In noisy environments, the microphone also captures the ambient noise in addition to the useful signal. In consequence, mobile terminals are generally equipped with speech signal processing algorithms in order to maintain and guarantee acceptable speech quality to the users. In this paper we will focus on the echo problem.

Most approaches to single-channel (SC) acoustic echo cancellation consist of an adaptive filter followed by an echo postfilter [1]. Nevertheless, the performance of SC echo cancellation approaches are sometimes unsatisfactory and a trade-off between echo suppression during echo-only periods and distortion of near-end speech introduced by the postfilter during double-talk periods [2]. Acoustic echo cancellation algorithms for SC terminals have received the most attention in last years. In this paper we propose an approach to dual channel (DC) echo cancellation.

This paper presents our recent work in echo cancellation for dual microphone mobile terminals (i.e equipped with one loudspeaker and two microphones). Given such an architecture, echo cancellation can still be achieved by adaptive filtering followed by a postfilter [3]. Adaptive echo cancellation can be achieved using one adaptive filter for each microphone path. As is the case for single microphone terminals, algorithms such as the classic normalized least mean square (NLMS) approach can be efficiently used. The contribution in this paper relates to the postfilter which uses two microphones signals instead of one as it is traditionally the case. The proposed postfilter is later on assessed and compared to an existing SC postfilter [2].

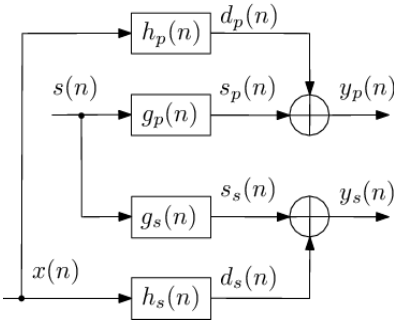


Fig. 1: Signal model

The remainder of this paper is organized as follows. In Section 2 we describe the echo problem in hands-free DC terminals. Section 3 presents the proposed echo processing scheme for dual microphone terminals. Experimental results are presented in Section 4 while conclusions are presented in Section 5.

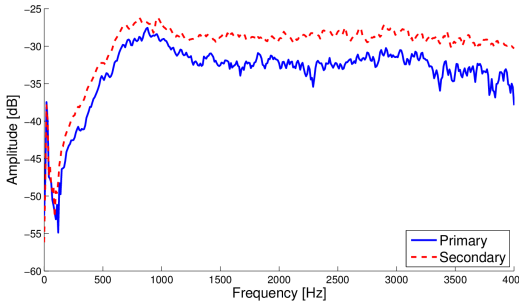
2. ECHO PROBLEM IN DUAL-CHANNEL TERMINALS

In this section, we describe the problem of echo in case of hands-free dual microphones mobile terminals. In Section 2.1 we propose a signal model for the echo problem in dual microphones mobile terminals. Section 2.2 compares the proposed signal model to data measured with a mock-up phone in realistic environments.

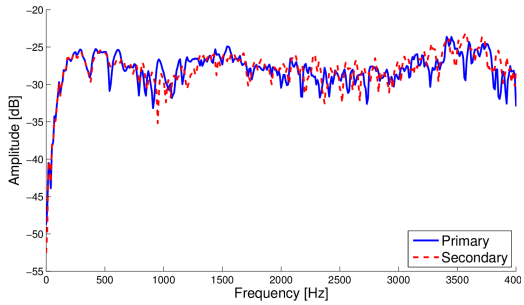
2.1. Signal model

We consider a mock-up mobile terminal equipped with one loudspeaker and two microphones. The microphones are placed at opposite corners of the phone as in [4]: one at the top corner and the other at the opposite bottom corner. The loudspeaker is placed at the back of the mock-up phone so as to simulate the loudspeaker used in mobile terminals in hands-free mode. The loudspeaker is placed so as to be slightly closer to the top microphone. Positioning the loudspeaker closer to one microphone corresponds to The bottom microphone observation is considered as being the primary observation and the top as being the secondary and are referred to as $y_p(n)$ and $y_s(n)$, respectively.

The signal model matching the physical interactions between the acoustic sources and the transducers of our system is showed



(a) Frequency responses between loudspeaker and microphones



(b) Frequency responses between artificial mouth and microphones

Fig. 2: Frequency responses with the phone placed in front of the artificial mouth

in Figure 1. The two microphones receive a modified version of the near-end speech signal $s(n)$ and of the far-end speech signal $x(n)$ played by the loudspeaker. Both the near-end speech and the far-end speech signals reflect in the near-end environment and before being received by each microphone. The acoustic paths between the near-end speech signal and the primary and secondary microphones are denoted $g_p(n)$ and $g_s(n)$ respectively while the coupling between the loudspeaker and each microphone is denoted by $h_p(n)$ and $h_s(n)$for the primary and secondary microphones respectively. The echo signals and the near-end speech signals at the primary or secondary microphones are denoted $d_u(n)$ and $s_u(n)$ with $u \in \{p, s\}$.

2.2. Analysis of recordings with hands-free terminals

In order to validate the signal model introduced in the above section, we performed some impulse responses measurements with the mock-up phone in different acoustic environments. A mannequin (HEAD Acoustics HMS II.3) with artificial mouth simulator is used to simulate the near-end speaker. We used two different phone positions: one where the phone is placed at a distance of $30cm$ directly in front of the artificial mouth and another where the phone is placed on a table according to ITU-T recommendations [5]. The recording are performed in different acoustic environments: office, meeting room. In all our recordings, the phone is placed so that the two microphones are approximately at equal distance to the artificial mouth.

Figure 2 (a) shows an example of frequency responses for the acoustic path between the loudspeaker and each microphone. The profiles show that the loudspeaker signal received by the microphones are not equally attenuated by the acoustic environment for each microphone. This shows the necessity to encounter for these differences by considering two acoustic echo paths in the signal

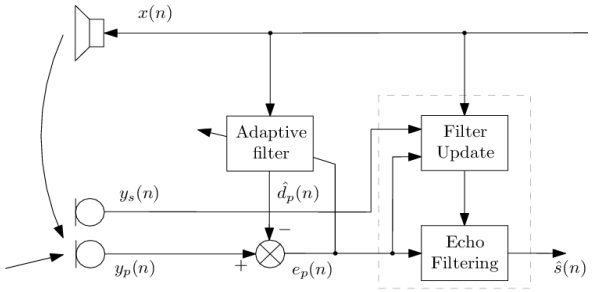


Fig. 3: Echo processing scheme

model.

Figure 2 (b) shows an example of frequency responses between the artificial mouth and the microphones. We see that both impulse responses are very similar. These similarities can be explained by the position of the microphones compared to the artificial mouth. For this reason, we assume in the following that $g_p(n) = g_s(n)$.

3. PROPOSED ECHO PROCESSING SCHEME

In this section we introduce the dual-channel (DC) echo processing scheme that we propose for the DC echo problem presented above. Similarly to single channel echo problem, the proposed scheme is composed of an adaptive echo canceler followed by an echo postfilter. Section 3.1 describes the proposed echo processing scheme while Section 3.2 presents the DC echo PSD estimate used within the postfilter.

3.1. Echo processing

The proposed DC echo processing scheme for the dual-microphone echo problem is illustrated in Figure 3 and is composed of an adaptive filter followed by an echo postfilter. The novelty in our system resides in the echo postfilter which uses two microphones paths.

Adaptive echo cancellation can be achieved by any standard adaptive filter such as normalized least square (NLMS) with a fixed or variable stepsize [3]. The use of the adaptive filtering here is justified by the fact the echo signal is generated by one loudspeaker therefore a standard adaptive filter can be used to achieve echo cancellation for each microphone path. To keep the computational complexity of the proposed DC echo processing scheme at a similar level to that of a SC scheme, we use only one adaptive filter placed on the primary microphone path as shown in Figure 3. The error signal $e_p(n)$ at the output of the adaptive filter can be written as:

$$\begin{aligned} e_p(n) &= y_p(n) - \hat{h}_p(n) * x(n) \\ e_p(n) &= g_p(n) * s(n) + \tilde{h}_p(n) * x(n) \\ e_p(n) &= s_p(n) + \tilde{d}_p(n) \end{aligned} \quad (1)$$

where $\hat{h}_p(n)$ is the estimate of $h_p(n)$ the echo path between the loudspeaker and primary microphone , $\tilde{d}_p(n)$ is the residual echo and $\tilde{h}_p(n) = h_p(n) - \hat{h}_p(n)$.

As illustrated in Figure 3, the postfilter is placed after the adaptive filter and is used to attenuate the residual echo at the output of the adaptive filter. We use a frequency domain echo postfilter since it gives good performance during double-talk periods [1]. The postfilter uses two microphones signals instead of one as it is often the case in the literature [1]. These microphone observations are used

together with the loudspeaker signal $x(n)$ to determine the residual echo suppression gains which we apply to the error signal $e_p(n)$ in the frequency domain to completely suppress the residual echo. Echo suppression itself is applied only to the primary microphone path. Accordingly existing echo suppression gain rules can be readily used. In our implementation, our postfilter gains are computed through a Wiener rule as follows

$$W(k,i) = \frac{SER(k,i)}{1+SER(k,i)}$$

where k is the frame index, i is the frequency index and the signal-to-echo ratio (SER) is estimated through decision directed approach [2, 6]. All like most gain rules, this gain rule requires only an estimate of the residual echo power spectrum density (PSD) $\hat{\Phi}^{\tilde{d}_p\tilde{d}_p}(k,i)$. For clarity, k and i indices will omitted in the remainder of this paper and will only be used when necessary.

3.2. Echo PSD estimate

In this section, we describe a mean of estimating $\hat{\Phi}^{\tilde{d}_p\tilde{d}_p}$ that uses the dual microphone signals. The residual echo PSD can be defined as:

$$\Phi^{\tilde{d}_p\tilde{d}_p} = |\tilde{H}_p|^2 \cdot \Phi^{xx}, \tag{2}$$

where Φ^{xx} is the PSD of $x(n)$ and $\tilde{H}_p$ is the frequency response of $\tilde{h}_p$.

Assuming $x(n)$ and $s(n)$ are uncorrelated, we can write the PSDs of $e_p(n)$ and $y_s(n)$ as:

$$\Phi^{e_pe_p} = |G_p|^2 \cdot \Phi^{ss} + |\tilde{H}_p|^2 \cdot \Phi^{xx} \tag{3}$$

$$\Phi^{y_sy_s} = |G_s|^2 \cdot \Phi^{ss} + |H_s|^2 \cdot \Phi^{xx} \tag{4}$$

where Φ^{ss} is the PSD of $s(n)$ and $G_{p|s}$ and H_s are the frequency response of $g_{p|s}$ and h_s respectively. By introducing the residual echo relative transfer functions (RTF) Γ defined as follows:

$$\Gamma = \frac{H_s}{\tilde{H}_p} \tag{5}$$

in Equation 4, we obtain:

$$\Phi^{y_sy_s} = |G_s|^2 \cdot \Phi^{ss} + |\Gamma|^2 \cdot |\tilde{H}_p|^2 \cdot \Phi^{xx} \tag{6}$$

By using the equality assumption between G_p and G_s justified in Section 2.2, we can derive an estimate of the residual echo PSD $\hat{\Phi}^{\tilde{d}_p\tilde{d}_p}$ from Equations 2, 3 and 6:

$$\hat{\Phi}^{\tilde{d}_p\tilde{d}_p} = \frac{\Phi^{e_pe_p} - \Phi^{y_sy_s}}{1-|\Gamma|^2}. \tag{7}$$

The PSDs $\Phi^{e_pe_p}$ and $\Phi^{y_sy_s}$ are computed through autoregressive smoothing. Γ can be obtained through the cross-PSDs between the loudspeaker signal $x(n)$ and $e_p(n)$ or $y_s(n)$. Assuming $x(n)$ and $s(n)$ are uncorrelated the cross-PSDs can be expressed as:

$$\Phi^{xe_p} = \tilde{H}_p \cdot \Phi^{xx} \quad \text{and} \quad \Phi^{xy_s} = H_s \cdot \Phi^{xx}, \tag{8}$$

From Equation 8, we deduce an estimate of the RTF $\hat{\Gamma}$:

$$\hat{\Gamma} = \frac{\Phi^{xy_s}}{\Phi^{xe_p}}. \tag{9}$$

We compute $\hat{\Gamma}$ during far-end speech activity periods. In our implementation, far-end activity periods are detected using a threshold on the loudspeaker signal energy.

4. EXPERIMENTS

In this section we assess the new dual-microphone residual echo PSD estimator by comparing its performance to an existing single microphone estimator from the literature. In Section 4.1 we describe the experimental setup used in our investigations. The proposed residual echo estimator accuracy is assessed in Section 4.2 and its performance for echo suppression performance are showed in Section 4.3.

4.1. Experimental setup

The different impulse responses recorded with our mock-up phone are used to generate a test database of speech signals. Microphone signals contains both echo-only and double-talk periods. The SER ranges from -5 to $10dB$ on the primary microphone.

Our DC echo processing scheme is compared to a SC echo processing scheme (i.e. SC adaptive filter followed by a postfilter) [2]. The SC setup uses only the primary microphone. The adaptive filter considered in our experiments is a NLMS adaptive filter [3] with variable stepsize. Both postfilters considered use the same gain rule (i.e. Wiener filter with SER estimated through decision directed approach [2, 6]). The DC and SC postfilters differ by the residual echo PSD estimator. The SC residual echo PSD estimator is that of [2] while the DC residual echo PSD estimator is that depicted in Section 3.2. In our experiments, the number of frequency bands M is set to 256 and the frame-by-frame conversion from time to frequency domain is done through short term Fourier transform with overlap add.

The assessment of the performance of the proposed postfilter is performed in two steps. The first part assesses the new PSD estimator accuracy by the mean of symmetric segmental logarithmic error [4] which can be expressed as follows:

$$logErr = \frac{1}{KM}\sum_{k}^{K}\sum_{i}^{M}\left|10\log_{10}\left[\frac{\Phi^{\tilde{d}_p\tilde{d}_p}(k,i)}{\hat{\Phi}^{\tilde{d}_p\tilde{d}_p}(k,i)}\right]\right| \tag{10}$$

where K is the number of frames and M is the number of subbands. The second part assesses the proposed DC PSD estimator echo suppression performance in terms of echo return loss enhancement (ERLE), of speech attenuation (SA), of cepstral distance (CD) and of informal listening tests. The ERLE reported here shows amount of echo suppression achieved by complete echo processing scheme (i.e. adaptive filtering and postfiltering) and is measured during echo-only periods. SA is used to measure the amount of speech attenuation introduced by the postfilter on the near-end speech signal during double-talk periods. SA is measured for the primary microphone signal as the attenuation between the clean speech $s_p(n)$ and the weighted speech signal $\bar{s}_p(n)$ [7] as follows:

$$SA = \frac{1}{\mathbb{K}}\sum_{\lambda}^{\mathbb{K}} 10\log_{10}\frac{\sum_{l=1}^{L} s_p^2(\lambda L+l)}{\sum_N \bar{s}_p^2(\lambda L+l)} \tag{11}$$

where L is length of the frames on which we compute the segmental SA and $\mathbb{K}$ represents the number of frames during which double talk occurs. Weighted speech signals $\bar{s}_p(n)$ are obtained according to [7]. The cepstral distance is measured similarly to [2] between $s_p(n)$ and $\bar{s}_p(n)$. Note that there is no need to assess the adaptive filtering part separately as we use the same adaptive filter for DC and SC echo processing.

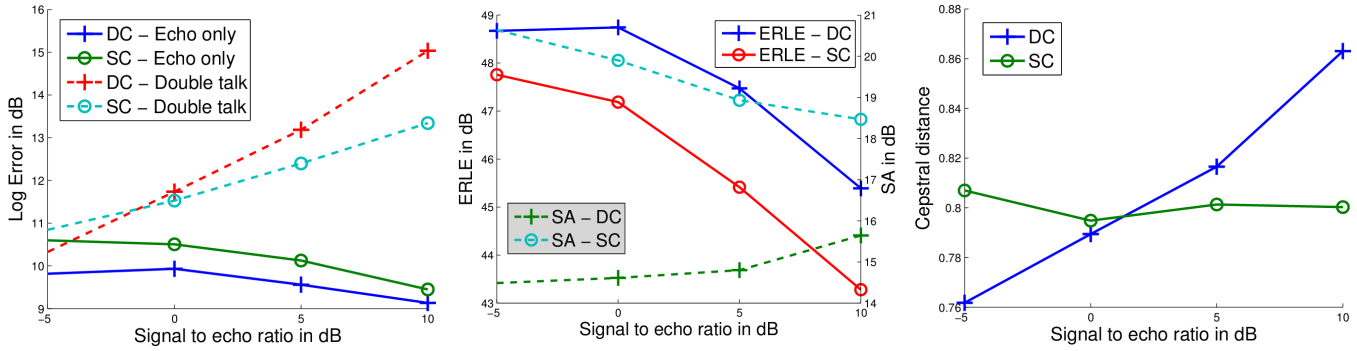


Fig. 4: Performance measure: (Left) PSD accuracy, (Middle) Average ERLE an SA, (Right) Cepstral distance

4.2. Residual echo PSD estimate assessment

Figure 4 (Left) illustrates the error in the residual echo PSD estimator during echo-only and double-talk periods. During echo-only periods the DC estimator slightly outperforms the SC estimator. We also observe that for both the DC and SC estimators the error decreases as the SER increases - albeit only slowly. The curves also show that, during double-talk periods, the error increases with the SER. This can be explained by the presence of near-end speech which disturbs the PSD estimation. Moreover, a high SER implies high near-end speech signal compared to echo (and therefore residual echo) and thus greater disturbance of the residual echo estimators. From Figure 4 (Left), we also observe that the DC estimator achieves better performance than the SC estimate for low SERs. In contrast, however, at high SERs ($SER > 0dB$), the SC estimator outperforms the DC estimators. The loss of performance of the DC can be justified by the fact that during double-talk, the presence of near-end disturbs the estimate of the RTF Γ as the cross-PSDs used for its computation do in practice contain a component dependent on the near-end speech signal. As the SER increases, the cross-PSDs component due to the near-end speech signal increase and so-doing leads to a wrong RTF estimate. Another reason which might explain the poor performance of the DC PSD estimator during double talk periods is the assumption according to which $g_p(n) = g_s(n)$ which does not really hold given the fact we use impulse responses measured in real environments.

4.3. Residual echo suppression

Figure 4 (Middle) shows the ERLE and SA curves. The ERLE curves shows that the DC echo postfilter achieves more echo suppression than the SC postfilter. This is a direct consequence of PSD estimate accuracy during echo-only periods. The SA curves show increasing attenuation of the near-end speech for the DC case with increasing with the SER while decreasing for the SC case. Such increase of the SA is undesirable as it means half-duplex situations. The DC postfilter nevertheless introduces less attenuation (up to 5dB) compared to the SC postfilter.

Figure 4 (Right) shows the CD during double talk periods. We note that at low SERs, the DC postfilter introduces less distortion than the SC postfilter while at higher SERs, the SC postfilter gives better performance. Moreover, the DC postfilter CD curve increases with the SER. This suggests that distortion introduced by the DC postfilter increases as the near-end speech signal becomes higher compared to echo (and residual echo). Our analysis is that this CD increase is a consequence of the SA increase observed in Figure 4 (Middle).

Informal listening tests show that the DC postfilter yields a slight better intelligibility of near-end speech compared to the SC postfilter during double-talk periods. The high SA introduced by the SC postfilter is perceptible and sometimes leads to complete suppression of the speech.

5. CONCLUSION

This paper reports a novel approach to acoustic echo cancellation for dual-microphones hands-free terminals. The proposed echo processing scheme is still composed an adaptive filter followed by an echo postfilter which exploits the dual-channel observation signals and requires an estimation of relative transfer function.

The new approach is assessed with data recorded with a mock-up phone in terms of objective performance metrics and informal listening. Our results shows that the proposed dual-channel echo postfilter offers significant advantages compared to existing single channel approaches, especially at low SERs. The main limitation of the new proposed echo cancellation scheme is that it is designed for hands-free scenarios and therefore have limited applications. Our future work aims to extend the proposed algorithm to work in handset mode.

6. REFERENCES

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

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

[3] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[4] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual microphone mobile phones exploiting power level differences," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 1693–1696.

[5] ITU-T, "ITU-T recommendation P.340: Transmission characteristics of handsfree telephones," 1996.

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

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