A CENTRALIZED ACOUSTIC ECHO CANCELLER EXPLOITING MASKING PROPERTIES OF THE HUMAN EAR

Xiaojian Lu and Benoît Champagne

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, H3A 2A7, Canada
{xlu,champagne}@tsp.ece.mcgill.ca

ABSTRACT

The design of shared, centralized acoustic echo canceller (AEC) for use in modern digital communication networks faces new challenges. Specially, the performance of conventional approaches for AEC is severely degraded by vocoder non-linearities along the transmission chain. In this paper, based on the analysis of the nonlinear echo path and the performance of the Wiener-type post-filter in the presence of vocoders, we propose a centralized AEC which incorporates a psychoacoustic post-filter. Computer experiments show that the proposed AEC is very promising for practical use in terms of high acoustic echo suppression, robust performance and ease of implementation.

1. INTRODUCTION

In traditional studies and applications of acoustic echo cancellers (AEC), it is commonly assumed that the AEC lies in a local hands-free terminal. In recent years, the use of shared, centralized AEC located at specific nodes along a digital communication networks has become very attractive for the industry. Indeed, the use of centralized AEC may not only significantly reduce the cost of the whole communication system, but also remarkably simplify the implementation of the local hands-free terminals. The latter is especially important for mobile wireless communications, where the power supply and the computational complexity are critical issues.

Nowadays, vocoders are increasingly used to reduce the speech transmission rate in modern digital networks, and especially in wireless communications and voice over IP applications. Because these vocoders exhibit nonlinear characteristics and cascade with the acoustic echo path, the entire echo path in such applications presents strong non-linearities. The performance of conventional AEC schemes, which mainly employ adaptive linear filter structures to model the acoustic echo path, is significantly degraded by these non-linearities [1].

The Wiener post-filtering technique and its variants, incorporating perceptual models of the human auditory system, have been widely used in speech enhancement, e.g. [2]. Recently, this hybrid technique was also used in a combined AEC-noise reduction approach to further suppress the residual echo resulting from the use of a shorter adaptive filter [3]. So far, most post-filters for AEC have been applied to, and analyzed in the context of linear acoustic echo path models [3, 4].

This work was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

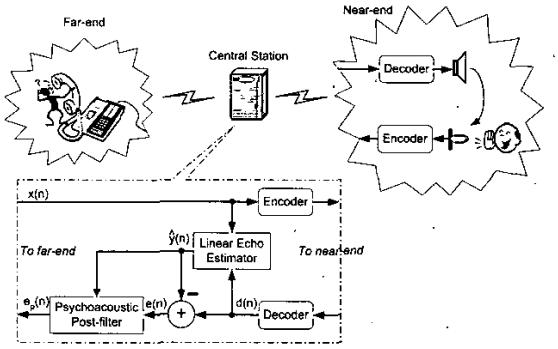

Fig. 1. Proposed centralized acoustic echo canceller

In this work, based on the analysis of the nonlinear echo path and the performance of the Wiener-type post-filter in the presence of vocoders, we propose a combined centralized AEC scheme which incorporates a robust post-filter that exploits the masking properties of the human ear. The proposed AEC, whose basic structure is illustrated in Fig. 1, is very promising for practical use in modern digital networks due to its high acoustic echo suppression, robust performance and ease of implementation, as demonstrated by computer experiments.

2. EFFECTS OF VOCODERS ON AEC

Vocoders such as G.729 [5] severely distort the speech signal in terms of waveform reproduction, but the perceptual distortion is limited. To analyze the properties of these vocoders from a rigorous mathematical standpoint is extremely difficult. Based on simplified considerations of the vocoder structure, combined with experimental observations, we may find that the vocoder has a complex nonlinear character. Furthermore the degree or extent of nonlinearity is signal dependent. For example, it can be verified that the degree of non-linearity is stronger for fricatives than it is for vowels.

In a centralized AEC application, the entire echo path consists of the acoustic echo path and electric signal paths comprising vocoders. The acoustic echo path is a characteristic of the loudspeaker-enclosure-microphone (LEM) system, which can be modelled as a linear system, albeit time-varying one. However, as pointed out above, the electric path with vocoders is a complex nonlinear system. Consequently, the complete echo path presents

a strong nonlinear character.

Referring to Fig. 1, $d(n)$ and $x(n)$ are the signals from the near-end and the far-end, respectively. Without loss generality, we assume that a hands-free device is only used in the near-end, so that $d(n)$ can be written as

$$d(n) = \nu(n) + y(n) \tag{1}$$

where $\nu(n)$ and $y(n)$ denote the near-end speech and the acoustic echo, respectively. The estimated echo, denoted by $\hat{y}(n)$, is then subtracted from $d(n)$, resulting in the residual signal $e(n)$

$$e(n) = \nu(n) + \delta(n) \tag{2}$$

where, $\delta(n) = y(n) - \hat{y}(n)$ is the residual echo.

Our experimental observations of the behaviour of conventional AECs operated in the nonlinear channel setting of Fig. 1 reveal the following: (1) a better tracking capability of the adaptive filter leads to a lower power of the residual echo signal $\delta(n)$; (2) the power of the residual echo signal $\delta(n)$ may exceed that of the echo signal $y(n)$ if the adaptation of the filter coefficients is frozen. Note that with conventional AECs, freezing the filter weight adaptation is a common way of avoiding possible divergence of the adaptive filter in the double-talk situation.

Based on these observations, we propose the use of the adaptive cross-spectral (ACS) algorithm [6], with suitable modifications as in [7], in the application of centralized AEC over nonlinear vocoder channels. This algorithm exploits the signal spectral correlations to estimate the acoustic echo path, resulting in a robust behaviour even in the presence of strong local disturbance signals. Since ACS keeps on adapting at all time, it does not require a double-talk detector. As a result, problem (2) above is avoided and the implementation of the AEC is considerably simplified.

Due to the nonlinear effects of the vocoder, the level of the residual echo $\delta(n)$ with ACS, and other conventional adaptive filtering algorithms for that matter, is not low enough to be imperceptible. Hence, further echo suppression is required in this application.

3. RESIDUAL ECHO SUPPRESSION

As pointed out in Section 2, the acoustic echo is not sufficiently cancelled by an adaptive filter, because the conventional approach of linear system identification cannot model the complex nonlinear system where vocoders are cascaded. However, a post-filter designed with the aim of preserving the useful signal, i.e. the near-end speech, as much as possible, while suppressing unwanted signal component, i.e. the residual echo, may be used to attenuate the residual acoustic echo.

3.1. Wiener-type post-filter in the presence of vocoders

Let $V(k;m)$ and $E(k;m)$ respectively denote the DFT of $\nu(n)$ and $e(n)$, where $k = 1, \ldots, K$ is the index of the frequency bins, and $m = 1, 2, \ldots$ is the frame index in the time domain. Define a linear estimator of $V(k;m)$ as

$$\hat{V}(k;m) = H(k;m)E(k;m) \tag{3}$$

where $H(k;m)$ is a real-valued frequency weighting function. Using a similar approach as in [8], it is easy to show that the optimal estimator is obtained as

$$H_{opt}(k;m) = \frac{S_{\nu\nu}(k;m)}{S_{\nu\nu}(k;m) + \alpha S_{\delta\delta}(k;m)} \tag{4}$$

where, $S_{\nu\nu}(k;m)$ and $S_{\delta\delta}(k;m)$ respectively denote the power spectral density (PSD) of $\nu(n)$ and $\delta(n)$; α is a positive constant.

The difficulty in using the post-filter of Eq. (4) is that both $\nu(n)$ and $\delta(n)$ are not directly measurable in a practical AEC system. However, referring to Eqs. (1)-(2), and assuming that $\nu(n)$, $y(n)$ and $\delta(n)$ are mutually uncorrelated, we can show that $S_{\nu\nu}(k;m) = S_{ed}(k;m)$. Thus, only the problem of estimating $S_{\delta\delta}(k;m)$ remains in Eq. (4).

In Fig. 2, the average magnitude spectra of $\delta(n)$ and $\hat{y}(n)$ are compared for two different speech frames, where the modified ACS algorithm was used. We can see that in both cases, the shapes of the magnitude spectra are alike. A possible interpretation of this phenomenon is provided below.

Nonlinearity usually brings in new frequency components. Because the spectrum of speech almost occupies the whole 4kHz bandwidth, the contribution of the system nonlinearity over that range is just to distort the spectrum by increasing or decreasing the existing frequency components. According to the properties of the vocoder, the distortion of the spectrum, especially in the regions of higher energy, is not supposed to be too severe so as to affect the auditory perception. The distortion also varies with acoustic phonetics: less distortion in the vowel-dominated frame (see Fig. 2(a)) than in the fricative-dominated frame (see Fig. 2(b)).

Based on these considerations, it appears reasonable to approximate $S_{\delta\delta}(k;m)$ by a scaled version of $S_{\hat{y}\hat{y}}(k;m)$. Hence, the post-filter in Eq. (4) becomes

$$H(k;m) = \frac{S_{ed}(k;m)}{S_{ed}(k;m) + \mu S_{\hat{y}\hat{y}}(k;m)} \tag{5}$$

where the attenuation factor, denoted by $\mu > 0$, controls the amount of echo suppression as well as the distortion of the near-end speech during the double-talk period. Due to the fact that the performance of the adaptive filter is severely degraded by the vocoder, the value of μ should be large enough to aggressively suppress the residual echo. We note that the PSD such as $S_{ed}(k;m)$ and $S_{\hat{y}\hat{y}}(k;m)$ in Eq. (5) can be estimated recursively [4].

Experimental results in Section 4 indicate that the acoustic echo is remarkably attenuated by the post-filter in Eq. (5). However, an annoying musical noise can be heard in the echo-suppressed signal. This situation with the use of Wiener-type post-filter is similar to the case of the linear acoustic echo path in [3].

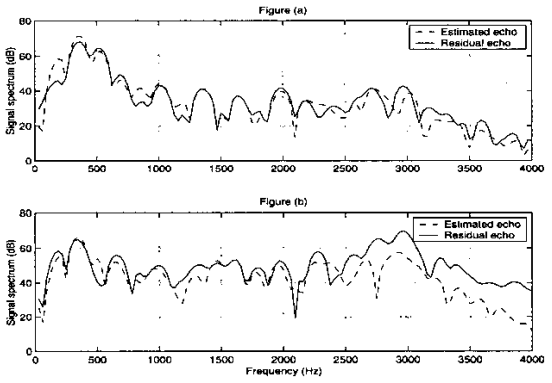

Fig. 2. Signal spectra of estimated echo $\hat{y}(n)$ and residual echo $\delta(n)$: (a) vowel-dominated frame, (b) fricative-dominated frame.

To mitigate these annoyances, a perceptual post-filter is proposed in the next section. By utilizing the masking properties of the human auditory system, the near-end speech signal should remain largely undistorted.

3.2. Post-filtering exploiting masking properties

Masking is a very important phenomenon of human auditory system where the perception of one sound is obscured by the perception of another [9]. The masking threshold is a level below which a weak signal becomes inaudible in the presence of a strong signal.

Accordingly, we only need to suppress the part of the residual echo with higher PSD until it is below the masking threshold. Given the masking threshold $T(k;m)$, the gain of a psychoacoustic post-filter is [3]

$$G(k;m) = \min\left(\sqrt{\frac{T(k;m)}{S_{\delta\delta}(k;m)}}, 1\right) \tag{6}$$

The echo-suppressed signal is thus obtained

$$E_p(k;m) = G(k;m)E(k;m) \tag{7}$$

Apparently, the performance of the psychoacoustic post-filter depends on the estimation accuracy of $T(k;m)$ and $S_{\delta\delta}(k;m)$. Based on the characteristics of the nonlinear system where vocoders are present, we propose a psychoacoustic post-filter scheme where $T(k;m)$ and $S_{\delta\delta}(k;m)$ are reliably estimated in this nonlinear channel.

a) Calculation of the masking threshold

Ideally, $T(k;m)$ is calculated from $\nu(n)$ which, however, is unmeasurable in this application. In practice, the estimated near-end speech, denoted by $\hat{\nu}(n)$, can be obtained from Eqs. (3) and (5). Indeed, this suboptimal estimator of $\nu(n)$, i.e. the post-filter shown in (5), minimizes the distortion of $\nu(n)$ while significantly attenuates $\delta(n)$. As pointed out in Section 3.1, the musical noise, which is caused by discontinuous spectral peaks, appears in $\hat{\nu}(n)$. However, it does not notably affect the calculation of $T(k;m)$ due to the properties of the auditory model, where these peaks are smoothed by spectral (and temporal) averaging. Hence, through the perceptual model, e.g. Johnston model [10], the masking threshold $\hat{T}(k;m)$ is computed with the input of $\hat{\nu}(n)$.

b) Estimation of the PSD of the residual echo

Due to the nonlinearity of the echo path, the estimation of the PSD of the residual echo by conventional means for linear systems [3] is severely degraded. Instead, we use the method of power spectral subtraction to find an approximate $S_{\delta\delta}(k;m)$.

Exploiting the signal relation in Eq. (2), we have

$$S_{ee}(k;m) = S_{\nu\nu}(k;m) + S_{\delta\delta}(k;m) \tag{8}$$

Replacing $S_{\nu\nu}(k;m)$ by its estimate, $\hat{S}_{\nu\nu}(k;m)$, and considering that the PSD is nonnegative, resulting in

$$\hat{S}_{\delta\delta}(k;m) = \max\{[S_{ee}(k;m) - \hat{S}_{\nu\nu}(k;m)], 0\} \tag{9}$$

c) Reduction of the masking distortion

The perceptual post-filter produces spectral (and temporal) averaging to smooth the residual signal, which effectively reduces the saliency of musical noise. Unfortunately, this may also impair the intelligibility of the near-end speech, because both $\hat{T}(k;m)$ and $\hat{S}_{\delta\delta}(k;m)$ are computed based on $\hat{\nu}(n)$ which was sometimes over-attenuated in order to aggressively suppress the residual echo.

To reduce this masking distortion, the proposed psychoacoustic post-filter stops attenuating the residual echo a few decibels above the masking threshold. Consequently, the signal distortion is reduced so that the intelligibility is improved. The post-filtering gain is then modified as

$$G(k;m) = \min\left(\sqrt{\frac{10^{P/10}\hat{T}(k;m)}{\hat{S}_{\delta\delta}(k;m)}}, 1\right) \tag{10}$$

where P (unit:dB) is the relaxation factor which may be chosen between $0 \sim 10$dB.

4. EXPERIMENTAL RESULTS

Computer experiments were conducted based on the platform illustrated in Fig. 1, where the vocoders were G.729. Real speech was used as the testing signals. The LEM system of the platform was simulated to represent the cab of a vehicle whose impulse response was about 40ms long.

The frame length of the psychoacoustic post-filter was 128 samples, with 50% overlap, where Hanning window was applied. Accordingly, the maximum delay of this AEC was about 16ms. Other parameters of the post-filter were set as $\mu = 40$ and $P = 5$dB.

At the beginning of our experiments, we tested the effects of vocoders on the conventional AEC, e.g. the modified ACS algorithm [7], as well as the performance of the post-filters in the presence of vocoders. The near-end speech was absent in this case, and the echo return loss enhancement (ERLE) was used to evaluate the echo suppression, which is defined by

$$\text{ERLE} = 10\log_{10}\frac{\sum_n y^2(n)}{\sum_n \delta^2(n)} \tag{11}$$

Similar to other conventional adaptive filtering algorithms, the performance of the modified ACS algorithm, shown in Fig. 3(a), is significantly degraded when vocoders are present along the echo

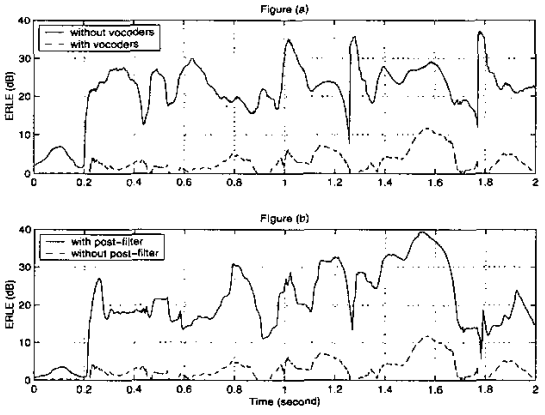

Fig. 3. (a) effect of the vocoder on the conventional AEC, (b) performance of Wiener-type post-filter.

path. Fig. 3(b) exhibits that the post-filter of Eq. (5) can aggressively attenuate the residual acoustic echo so that the degradation caused by vocoders can be compensated. We note that the psychoacoustic post-filter has very similar performance in terms of ERLE, compared to the Wiener-type post-filter.

Then a more realistic scenario was simulated, where three situations were considered, namely, far-end single-talk, double-talk and near-end single talk. Fig. 4 displays the waveforms of the signals in these situations. The near-end signal $d(n)$ that consists of $y(n)$ and $\nu(n)$ is shown in Fig. 4(a). We note that the real $\nu(n)$ cannot be obtained in the double-talk situation when $y(n)$ and $\nu(n)$ are mixed, due to the non-linearities of the vocoder. However, $d(n)$ is a reasonable approximation of $\nu(n)$ when the acoustic echo is absent, which is plotted in Fig. 4(d).

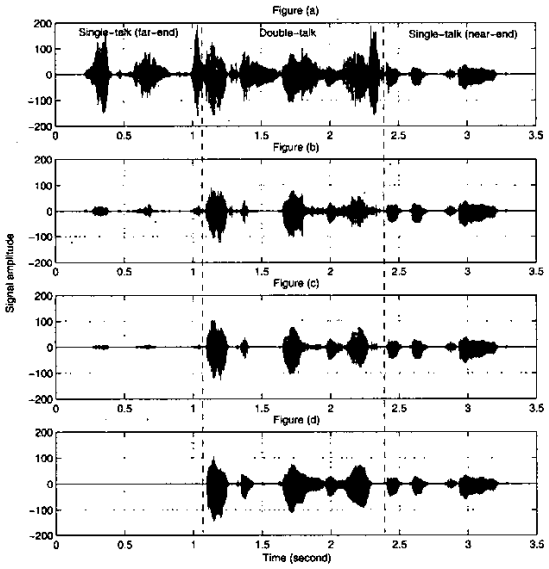

Fig. 4. Waveforms of signals: (a) near-end signal $d(n)$, (b) echo-suppressed signal $e_p^r(n)$, (c) echo-suppressed signal $e_p(n)$, (d) approximate near-end speech $\nu(n)$.

The echo-suppressed signals $e_p(n)$, shown in Fig. 4(c), is the output of the proposed AEC. Because the estimation of the residual echo plays the key role in the performance of the psychoacoustic post-filter, $S_{\delta\delta}(k;m)$ was also estimated by the approach in [3] for comparison. This led to a different echo-suppressed signal $e_p^r(n)$ which is displayed in Fig. 4(b). Comparing $e_p(n)$ and $e_p^r(n)$ with the approximate $\nu(n)$ in Fig. 4, one can find that, in the presence of vocoders, the proposed AEC has advantages in terms of higher echo suppression and of less distortion to the near-end speech during the double-talk period.

Furthermore, informal listening tests were also conducted. The Wiener-type post-filter produced strong musical noise, although it remarkably suppressed the residual echo. On the contrary, the musical noise produced by the proposed psychoacoustic post-filter was almost imperceptible. The listening comparison between $e_p(n)$ and $e_p^r(n)$ was also done. It was found that $e_p(n)$ has better quality in terms of less residual echo, less musical noise and less distortion to the near-end speech during the double-talk period.

We note that the proposed AEC still brought somewhat perceptual distortion to the near-end speech, although it did not affect the intelligibility. However, in the practical scenario, this may not be a critical issue since the far-end user is not sensitive to the speech quality of the near-end user when he/she is speaking.

In this paper, only the results with the Johnston perceptual model [10] have been shown. However, some sophisticated perceptual models, e.g. PEAQ [11], which consider the temporal masking as well as the spectral masking were also tested in our research. We note that no obvious improvement has been observed for the advanced models in this application. Therefore, the Johnston model is more suitable for the AEC since it has a simpler structure and lower computational complexity.

5. CONCLUSION

We have presented a robust centralized AEC which combines a psychoacoustic post-filter scheme with the modified ACS algorithm for the use of echo suppression over a nonlinear channel where vocoders are present along the echo path. The experimental results show that the proposed AEC is very promising for practical use in terms of remarkable echo suppression, robust behaviour and ease of implementation. Moreover, this AEC does not need the double-talk detection.

6. REFERENCES

[1] Y. Huang and R.A. Goubran, "Effects of vocoder distortion on network echo cancellation," in *Proc. ICME'00*, Aug. 2000, vol. 1, pp. 437–439.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[3] S. Gustafsson, R. Martin, P. Jex, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, Jul. 2002.

[4] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. ICASSP'97*, Apr. 1997, pp. 307–310.

[5] *ITU-T Recommendation G.729*, Mar. 1996.

[6] T. Okuno, M. Fukushima, and M. Tohyama, "Adaptive cross-spectral technique for acoustic echo cancellation," *IEICE Trans. Fundamentals*, vol. E82-A, no. 4, pp. 634–639, Apr. 1999.

[7] X. Lu and B. Champagne, "Acoustic echo cancellation over a non-linear channel," in *Proc. IWAENC'01*, Sep. 2001, pp. 139–142.

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[9] E. Zwicker and H. Fastl, *Psychoacoustics–Facts and Models*, Springer, New York, 2nd edition, 1999.

[10] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[11] *ITU-R Recommendation BS.1387-1*, Nov. 2001.