STFT-BASED MULTI-CHANNEL ACOUSTIC INTERFERENCE SUPPRESSOR

Carlos Avendano and Guillermo Garcia

Creative Advanced Technology Center
1500 Green Hills Road, Scotts Valley, CA 95067
{carlosa,guille}@atc.creative.com

ABSTRACT

In this paper we describe a system that suppresses the acoustic interference due to the coupling between the microphone and the loudspeakers of a hands-free multi-channel desktop audio system. The proposed system operates in the Short-Time Fourier Transform domain and uses spectral subtraction to suppress the unwanted interference, which consists of the local audio and the remote speech signal (echo). The interference estimate is obtained with a sub-band RLS-based adaptive multi-channel echo canceller. Test results show that under some adverse conditions and with low complexity constraints the system can achieve better and more consistent speech quality than a time-domain acoustic echo canceller.

1. INTRODUCTION

Voice over IP (VoIP) applications such as voice chat, teleconferencing, long distance telephony and network game playing have increasingly made the desktop audio system an important voice communication device. However, there are still several technical problems that need to be solved before VoIP systems can achieve a robust performance and deliver high-quality speech. Among those problems, the one that we address in this paper is the reduction of the interference due to the acoustic coupling between the microphone and the loudspeakers during hands-free operation. The setup considered represents a typical desktop system and consists of two (or more) users who are communicating over a data link, using a full-duplex multi-speaker single-microphone audio system (see Fig. 1). When the loudspeaker signals include the voice of the remote user $v(n)$, it will be transmitted back to the remote site along with the desired speech signal and will be perceived as an echo. Another source of interference arises when the local user is simultaneously playing music or other audio material through the loudspeakers.

One solution to this problem is the traditional acoustic echo canceller (AEC). The idea is to model the impulse responses between loudspeakers and microphone $h_L(n)$ and $h_R(n)$, and to filter the known loudspeaker outputs with these filters to generate an estimate of the interfering signal. Although the AEC is a straightforward technique, there are various problems associated with it. Since the cancellation is performed directly on the waveforms, the algorithm is very sensitive to the misalignment in the room transfer function estimates [1]. Another typical problem is that the room impulse responses are hundreds of milliseconds long, thus the adaptive filters that serve to model these responses need to be long (thousands of coefficients), making the algorithm complexity large and the convergence slow.

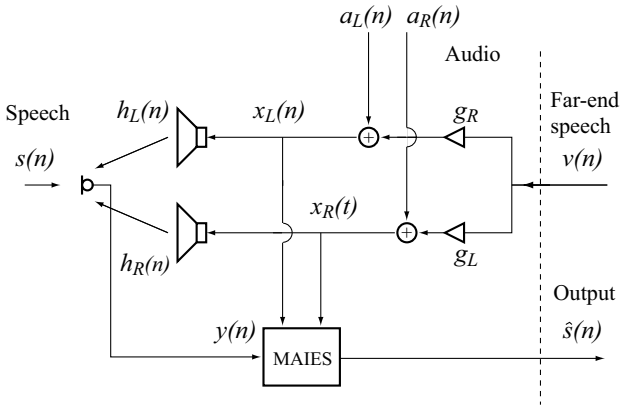

Fig. 1. *Desktop audio system used for VoIP. The acoustic interference is due to the panned monaural far-end speech $v(n)$ and the stereo local audio $a(n)$.*

2. PROPOSED SYSTEM

The system proposed in this paper is a multi-channel acoustic interference and echo suppressor (MAIES) that operates in the frequency domain. In contrast to the traditional AEC, where the goal is to cancel the interference at the waveform level, the MAIES suppresses the interference in the magnitude of the Short-Time Fourier Transform (STFT) via spectral subtraction [2]. To estimate the interference component, the system uses a sub-band adaptive AEC that operates on the STFT trajectories of the reference signals $x_L(n)$ and $x_R(n)$. After subtraction, the clean speech is synthesized using the new short-time magnitude estimate and the short-time phase of the original microphone input. The block diagram of the MAIES is shown in Fig. 2.

As we show later, there are various advantages in adopting this strategy. Spectral subtraction allows us to control the amount of interference that is suppressed (in exchange for some speech distortion), thus accurate estimation of the room responses is not essential. This flexibility also results in robustness of the algorithm to handle abrupt changes in the environment. Additionally, the computational complexity is significantly reduced compared to time-domain echo cancellation systems.

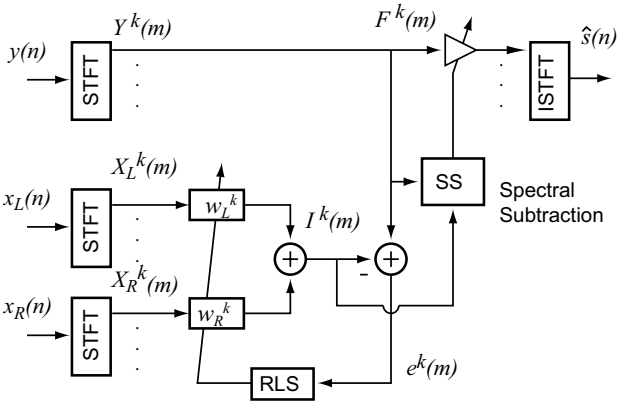

Fig. 2. *MAIES block diagram. Only the k^{th} sub-band trajectory is shown.*

2.1. Interference Suppression

Without loss of generality, consider the two-loudspeaker system of Fig. 1. We observe that the microphone input $y(n)$ can be written in terms of the desired (target) speech signal $s(n)$, the loudspeaker signals $x_L(n)$ and $x_R(n)$ and the room impulse responses [1] $h_L(n)$ and $h_R(n)$ as

$$y(n) = x_L(n) * h_L(n) + x_R(n) * h_R(n) + s(n), \quad (1)$$

where $*$ denotes convolution. Taking the discrete-time STFT of $y(n)$ yields the following equation:

$$Y(m,k) = I(m,k) + S(m,k), \quad (2)$$

with

$$I(m,k) = X_L(m,k) * H_L^k(m) + X_R(m,k) * H_R^k(m),$$

where $*$ now denotes convolution along the time dimension m, k is the discrete frequency index, $Y(m,k)$, $X_L(m,k)$, $X_R(m,k)$ and $S(m,k)$ are the STFTs of $y(n)$, $x_L(n)$, $x_R(n)$ and $s(n)$ respectively, and $H_L^k(m)$ and $H_R^k(m)$ are frequency dependent filters that represent the contribution at frequency k of the room impulse responses $h_L(n)$ and $h_R(n)$ in the STFT domain (notice that these terms act as filters along the time axis of the STFT [3]). The spectral subtraction algorithm consists of estimating the short-time magnitude spectrum of the interference and subtracting it from the magnitude of $Y(m,k)$. In a more general form of spectral subtraction, the estimate of the clean speech spectrum can be computed as:

$$|\hat{S}(m,k)| = [|Y(m,k)|^\alpha - \beta|\hat{I}(m,k)|^\alpha]^{\frac{1}{\alpha}}, \quad (3)$$

where the parameters α and β serve to control the amount of subtraction, and represent a compromise between interference attenuation and signal distortion [4]. Notice that this is a deterministic formulation of the spectral subtraction technique, as opposed to the stochastic form used to handle additive Gaussian noise [5].The interference estimate is obtained as

$$\hat{I}(m,k) = X_L(m,k) * W_L^k(m) + X_R(m,k) * W_R^k(m), \quad (4)$$

where $W_L^k(m)$ and $W_R^k(m)$ are estimates of $H_L^k(m)$ and $H_L^k(m)$ respectively.

Since we are avoiding the problem of phase estimation, the best clean speech estimate that we can obtain will have to use the short-time phase of the microphone input, i.e.

$$\hat{S}(m,k) = |\hat{S}(m,k)|e^{j\angle Y(m,k)} \quad (5)$$

This estimate is referred to as the theoretical limit in STFT estimation, and based on the threshold of phase perception, it has been shown that as long as the signal-to-noise ratio (SNR) is larger than about 5 dB, the use of the original "noisy" phase will not cause perceptible distortion of the reconstructed speech signal [5]. It can also be shown that the subtraction in (3) is essentially a time-varying multiplicative operation where the modification depends on the relative levels of the desired signal and the interference [4], and (5) can be written as

$$\hat{S}(m,k) = Y(m,k)F^k(m), \quad (6)$$

where

$$F^k(m) = \left[1 - \frac{\beta|\hat{I}(m,k)|^\alpha}{|Y(m,k)|^\alpha}\right]^{\frac{1}{\alpha}}.$$

Since we are using the original short-time phase for synthesis, the spectral subtraction operation can then be interpreted as applying a time-varying real-valued gain factor to the STFT trajectories (see Fig. 2).

2.2. Interference Estimator

The estimate of the interference $\hat{I}(m,k)$ may in principle be obtained in the time domain, such as AEC systems do. However, since we operate in the STFT domain it is advantageous and less computationally expensive (see section 3.1) to characterize and model the effect of the room impulse responses directly in this domain and use equation(4) to obtain the interference component.

To obtain the room response estimates $W_L^k(m)$ and $W_R^k(m)$ we invoke the filter-bank interpretation of the STFT and notice that at each center frequency k, the output of the STFT is a complex time series. During intervals where the target signal is not present ($s(n) = 0$) we can use equations (2) and (4) to write the error for the k^{th} time trajectory as $e^k(m) = Y(m,k) - \hat{I}(m,k)$.

If we let the estimates $W_L^k(m)$ and $W_R^k(m)$ be FIR filters and if we assume that the loudspeaker signals are stationary and uncorrelated, minimization of $e^k(m)$ in the mean squared sense will lead to the optimal Wiener solution. However, since the environment and the loudspeaker signals are not stationary, we minimize this error recursively using the RLS algorithm. Notice that we are minimizing the error for each time trajectory separately, thus we are assuming that the sub-bands are independent. This is in contrast to other sub-band systems where the cost function includes the sum of sub-band errors [6]. While our strategy does not guarantee that the misadjustment in the time domain will be minimum, we are only interested in obtaining the interference estimate in the

[1] The impulse responses $h_L(n)$ and $h_R(n)$ include the room acoustics as well as the response of the transducers and other linear distortions introduced by the audio equipment.

STFT domain. As we discuss later, the sub-band independence assumption represents a reduction in complexity that will result in a more efficient system.

2.3. Non-Uniqueness Problem

In practical situations, the correlation between loudspeaker signals may be large and the estimation of the room responses will suffer from the well-known non-uniqueness problem [1]. Several techniques have been proposed to overcome the non-uniqueness problem [1, 7, 8, 9]. Among those methods, non-linear shakers (half-wave rectification) seem to be the most effective for de-correlating band-limited speech signals [9]. However, in the context of the MAIES, non-linear distortion may be objectionable when the loudspeaker signals include high quality audio. A very moderate an imperceptible amplitude modulation (less than 0.5% modulation index) with very low carrier frequencies (32.25 Hz and 31.75 Hz for left and right channels respectively) close to the center frequency of the sub-band trajectories seems to provide enough de-correlation for practical purposes.

3. IMPLEMENTATION

The internal operation of the system was set to 12 kHz sampling rate. The reference signals, which may contain high quality audio, are down-sampled at the input of the system. The STFT analysis is performed every 10 ms ($J = 128$ samples) on 40 ms-long frames ($N = 512$ samples). The filters $W_L^k(m)$ and $W_R^k(m)$ were implemented as two complex 3-tap ($L = 6$) FIR filters. Given the window hop size of 10 ms, these filters have an effective length of 60 ms ($N + (L - 1) * J$ samples). Since the adaptive filters are short, the RLS algorithm is preferred over its fast implementations, which introduce instability problems (e.g. FRLS). The ranges of the spectral subtraction parameters that achieve the best perceptual quality are $0.5 \leq \alpha \leq 1$ and $1 \leq \beta \leq 1.2$.

A target signal detector based on short time spectral matching was implemented to stop the adaptation during local speech activity. The complete system has been implemented in real-time and can run as a host-based application on a 400 MHz personal computer.

3.1. Complexity

The number of real multiply-adds per output sample required by MAIES is $\frac{N}{2}\frac{1}{J}(14L^2 + 12L) + 40\frac{N}{2J} + 4\frac{3}{2}\frac{N}{J}\log_2(N)$, (for the adaptive algorithm, spectral subtraction and FFT's respectively) where N is the DFT length in samples, L is the total number of taps and J is the window hop size in samples. For a traditional time domain RLS-based stereo AEC the number of real multiply-adds per output sample is $4(L^2) + 3(L)$. Stabilized versions of the fast RLS algorithm have complexities that vary linearly with L, for example [10] reports a complexity of $14L$.

For the current implementation of MAIES, the effective length of the filter is 60 ms for each channel requiring $L = 6$ coefficients (two 3-tap filters per sub-band), thus the total cost is approximately 1496 multiply-adds per output sample. For the same effective length, the time-domain stereo AEC must adapt 1436 coefficients (two 768-tap filters), thus requiring about 9.5 million multiply-adds per output sample, and the fast RLS only $21,504$ multiply-adds per output sample. These quantities are respectively $6,300$ and 14 times larger than the cost of the MAIES.

4. EVALUATION

The system was evaluated in a series of simulations where its performance was measured in terms of the speech quality using the Itakura-Saito (IS) distance between the speech input and the speech estimate. This metric is commonly used to evaluate enhancement algorithms, where perceptual quality, rather than waveform fidelity is the requirement [11]. The IS distance will capture both the distortion due to spectral subtraction as well as the interference residual, thus it is also appropriate for evaluating a time domain AEC for comparison purposes.

4.1. Simulations

The speech data used consisted of eight seconds of speech from a male speaker. The loudspeaker signals included high quality stereo music and a center-panned monaural speech signal from another male speaker. The interference component was simulated convolving the loudspeaker signals with two room impulse responses measured in a real desktop environment. The measurements were windowed and truncated to include only the first 85 ms of the responses (4096 samples at 48 kHz sampling rate).

In AEC, when a change in the room responses occurs, some unwanted residual is heard at the output before the algorithm adapts to the new situation. The spectral subtraction operation in the MAIES provides the flexibility of controlling the amount of residual that is allowed at the output, in exchange for increased spectral distortion. Thus the MAIES is potentially more robust to misadjustment than the traditional AEC in this sense.

In the first simulation we evaluate the robustness to misadjustment of the system by perturbing the room responses and computing the average distortion over the entire signal. The level of the interference was set to a segmental signal-to-interference ration (SSIR) of 5 dB, and the spectral subtraction parameter values were both set to unity. After initial convergence, the adaptive algorithm was stopped and the room responses were perturbed by randomly adding or subtracting some fraction of their coefficient values. Tests in an office environment have shown that moderate subject motion may cause as much as 15% average change in the coefficients, while microphone, or loudspeaker relocation can cause much larger changes (greater than 50%). An AEC with the same effective filter length (30 ms) was also tested under these conditions and the resulting distortion values were compared.

The results are shown in Fig. 3(a), where it can be observed that while the MAIES has slightly poorer performance for small perturbations (an IS distance of less than 0.1 indicates no perceptible distortion [9]), it is more consistent than the more complex time domain canceller, whose performance degrades quickly as a function of perturbation. The dashed curve corresponds to the output of the MAIES when $\alpha = 0.75$, and indicates that the system output can be made more consistent by varying the spectral subtraction parameters, but at the cost of some additional distortion. We have observed that values of $\beta > 1$ result in reduced residual, but at the expense of more perceptible distortion.

From the first simulation we also observe that under no perturbation the MAIES introduces distortion. The amount of distortion

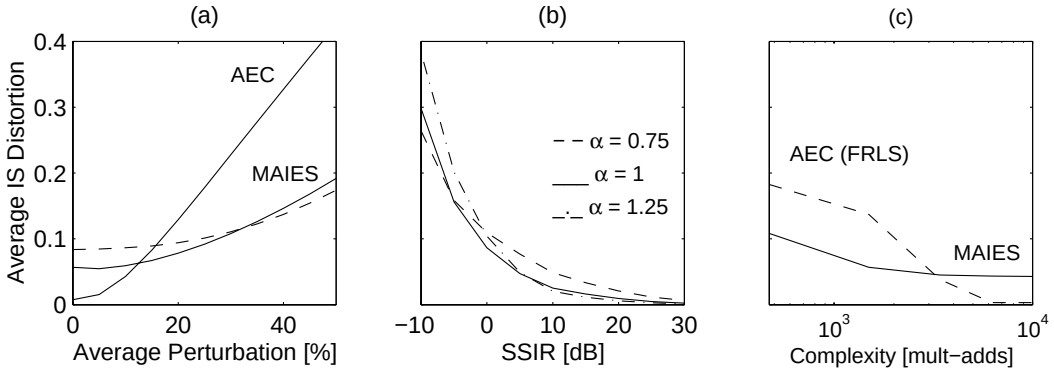

Fig. 3. *Simulation Results. (a) Average distortion versus acoustic channel perturbation, (b) distortion as a function of SSIR and (c) distortion versus complexity comparison.*

is proportional to the SSIR, thus the second simulation involved varying the SSIR by scaling the interference signals and computing the average distortion after the system converged. The system was evaluated for three settings of spectral subtraction parameters: $\alpha = 0.75, 1, 1.25$ and $\beta = 1$. The results are shown in Fig. 3(b), where we see that the distortion is below 0.1 for values of SSIR greater than 5 dB. We also see that the distortion is a function of α. For low SSIR, lower values of α are preferred and for high SSIR, higher values of α introduce less distortion.

The third simulation consisted of equalizing the complexity levels of the MAIES ($\alpha = 1$ and $\beta = 1$) and a time domain FRLS-based AEC, and comparing the distortion at the outputs for different levels of complexity. The SSIR level was set to 5 dB, and the complexity was varied in both systems by changing the effective length of the modeling filters. The result is shown in Fig. 3(c), where we observe that for low complexity, the MAIES is superior to time-domain AEC. For complexities greater than about 3500 multiply adds per sample, the AEC solution has better performance.

5. CONCLUSION

We have presented a low-complexity algorithm that suppresses unwanted acoustic interference and echo in multi-channel desktop audio systems. The use of spectral subtraction allows us to control the amount of interference suppression with a slight increase in speech distortion. The interference estimator uses a sub-band adaptive echo canceller, which is numerically stable and less complex than a time-domain AEC. The system is also more robust to changes of the acoustic environment or misalignment in the room transfer function. Additionally, processing the signals in the STFT domain is advantageous if short-time feature extraction or noise suppression algorithms follow down the signal-processing path of the speech signal.

6. ACKNOWLEDGEMENTS

The authors would like to thank Mark Dolson, Jean Laroche, and Brian Link for their useful suggestions and comments, and their encouragement to write this paper.

7. REFERENCES

[1] M.M. Sondhi, D.R. Morgan and J.L. Hall, "Stereophonic Acoustic Echo Cancellation -An Overview of the Fundamental Problem," IEEE Signal Processing Letters, Vol.2 No. 8., August 1995.

[2] S.F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE Trans ASSP, Vol. 29, pp. 113-120, April 1979.

[3] C. Avendano, "Temporal Processing of Speech in a Time-Feature Space," Ph.D. Thesis, Oregon Graduate Institute of Science & Technology, April 1997.

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," Proc. of the IEEE , Vol. 67, No. 12, pp. 1586-1604, December 1979.

[5] P. Vary, "Noise Suppression by Spectral Magnitude Estimation," Signal Processing, Vol. 8, pp. 387-400, 1985.

[6] W. Kellerman, "Analysis and Design of Multi-rate Systems for Cancellation of Acoustical Echoes," Proc. IEEE ICASSP'88, pp. 2570-2573, April 1988.

[7] S. Shimauchi, S. Makino, Y. Haneda, A. Nakagawa and S. Sakaucho, "A Stereo Echo Canceller Implemented Using a Stereo Shaker and a Duo-Filter Control System," Proc. IEEE ICASSP'99, Phoenix, 1999.

[8] S.G Sankaran and A.A. Beex, "Stereophonic Acoustic Echo Cancellation Using NLMS with Orthogonal Correction Factors," in Proc. Intl. Workshop on Acoustic Echo and Noise Control, Pocono Manor, PA, pp.40-43, September 1999.

[9] J. Benesty, D. R. Morgan, J. L. Hall, and M. M. Sondhi, "Synthesized stereo combined with acoustic echo cancellation for desktop conferencing," in Proc. IEEE ICASSP, 1999, vol. 2, pp. 853-856.

[10] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive Filtering Algorithms for Stereophonic Acoustic Echo Cancellation." Proc. IEEE ICASSP'95, pp. 3099-3102, Detroit, 1995.

[11] J.R. Deller, J.H.L. Hansen and J.G. Proakis, "Discrete-Time Processing of Speech Signals," IEEE Press, New York, 1993.