Proc. Biennial Symp. Commun. (Kingston, ON), pp. 324-327, June 1990

APPLICATION OF COMPLEX CEPSTRUM TO ACOUSTIC DEREVERBERATION

Duncan Bees [1], Peter Kabal [1,2], and Maier Blostein [1,2]

[1] Electrical Engineering
McGill University
Montreal, Quebec
Canada, H3A 2A7

[2] INRS-Télécommunications
Université du Québec
Verdun, Quebec
Canada, H3E 1H6

ABSTRACT

Speech in rooms is subject to degradation caused by acoustic reverberation. Signal processing techniques to remove reverberation have required multiple microphones or knowledge of the room impulse response. In this paper, complex cepstral deconvolution is applied to acoustic dereverberation. A new approach to the segmentation and windowing procedure for speech improves the complex cepstral identification of the reverberant impulse response, and least squares inverse filters are used to remove the estimated impulse response from the reverberant speech. Although complete removal of the impulse response is not possible, reduction of reverberation with this technique is demonstrated.

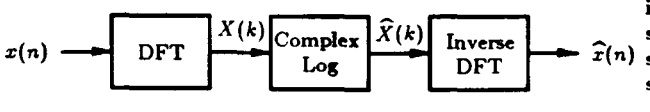
1. INTRODUCTION

Speech in rooms is subject to reverberation. Researchers have classified reverberations according to their perceptual effects: short delay time echoes modify the short time spectrum of speech and the resulting effect is called coloration; echoes with delays greater than perhaps 50 ms are perceived as audible copies of the speech. The durations and patterns of impulse responses in rooms vary with the geometry and reflective characteristics of the walls and other surfaces. Also, the amount of reverberant speech compared with direct speech increases with separation between talker and listener.

Speech dereverberation techniques have focused on processing the signals received at two or more microphones. Single microphone reverberant speech enhancement, applicable for example to hands-free telephony, typically requires knowledge of the room impulse response characteristics. Techniques for estimation of such characteristics from the reverberant speech alone have not been described. In this paper, the complex cepstrum is applied to speech dereverberation. The complex cepstrum is used to estimate the room impulse response for delays within several hundred milliseconds of the direct path speech. From the estimated impulse response, linear least square filters are designed with the goal of reducing the perceptual effects of the reverberation.

1.1 Complex Cepstrum

The complex cepstrum is described in [1]. It is a two-sided (non-causal), infinite sequence related to the time domain sequence by a non-linear transformation. For the discrete time signal $x(n)$, the characteristic system [1] by which the complex cepstrum is calculated is the following:



The complex cepstrum has several properties which make the technique a candidate for deconvolution. First, signals which are combined convolutionally in the time domain have complex cepstra which are combined additively. As a result, deconvolution is reduced to subtraction in the cepstrum. Second, the complex cepstrum is a measure of the "frequency" of variation (known as quefrency) in the log spectrum, and so signals which vary slowly in the log spectrum may be separated from quickly varying signals by windowing the complex cepstrum. Speech is usually considered to be primarily slowly varying in the log spectrum and has a complex cepstrum concentrated about the cepstral origin. Echoes which are delayed from the direct path speech can be represented by impulse responses which have cepstra concentrated farther from the cepstral origin.

1.2 Echo Removal with Complex Cepstrum

The techniques by which speech may be separated from a convolved impulse response representing a simple pattern of echoes are described in [2]. In these techniques, the complex cepstrum is calculated from segments of the reverberant speech by means of phase unwrapping, and the cepstral components corresponding to the impulse response are removed. If the complex cepstrum of the echoes are in the form of peaks, they are identified through a peak-picking procedure and the cepstral values at their locations are set to zero. Alternately, the calculated cepstrum is multiplied by a cepstral window function designed to preserve the speech cepstrum and remove the echo cepstrum. The remaining cepstrum is re-transformed to the time domain to form the enhanced speech.

We found these techniques, however, to be unsuitable for the dereverberation of speech subject to acoustic reverberation, where the impulse response of the echo is a much more complicated function, and one must deal with speech sequences of indefinite length. First, we found that the segmentation of the speech for the cepstral calcula n results in ral distortion in the form of noise and of crossover between causal and anti-causal cepstral terms; this is discussed later. Second, the phase unwrapping step [3] was found to fail occasionally, resulting in a loss of speech estimate for that segment. This effect is related to the presence of zeroes of the combined signal close to the unit circle. Third, the speech cepstrum is not ideally limited in range and the cepstral windowing procedure removes cepstral components corresponding to speech, resulting in distortion.

To overcome these problems, new approaches to estimating the room impulse response from the reverberant speech cepstrum were developed. The estimation of the echo impulse response was more successful than the direct estimation of the speech for several reasons. The impulse response was assumed

to be constant or to vary only slowly, and by forming the estimate of the impulse response from a moving average of cepstra from several speech segments, cepstral noise was attenuated and the speech cepstrum averaged towards zero. The speech segments themselves were defined using an "intelligent" method which reduced the segmentation error. The overall estimation procedure, which is summarized below, allows the identification of the impulse response with sufficient accuracy that a least squares filter designed to remove it achieved the desired effect of a reduction in reverberation.

2. DEREVERBERATION TECHNIQUE

In the dereverberation technique developed, the primary goal was to estimate accurately the reverberation impulse response. From this estimate, least squares filters were designed and applied to the reverberant speech. Figure 1 shows a block diagram of the dereverberation system.

Segmentation and windowing are the first steps performed and they are crucial to the correct estimation of the impulse response. Speech is a signal of indefinite length but it must be processed in finite length blocks. In breaking speech into segments the convolutional model between impulse response $h(n)$ and speech $s(n)$ is distorted. Each finite length segment $x_i(n)$ of the reverberant speech can be written [2] as

$$x_i(n) = s_i(n) * h(n) + e_i(n) \qquad (1)$$

where $e_i(n)$ is the segmentation error caused by the intrusion of the echo of $s_{i-1}(n)$ at the start of $x_i(n)$ and the truncation of the echo of $s_i(n)$ at the end of $x_i(n)$. The effect of $e_i(n)$ is to introduce distortion into the cepstrum which decreases the success of cepstral deconvolution.

The segmentation error may be reduced by multiplication by window functions with taper at the ends, but commonly used functions such as Hamming windows also distort the convolutional model. The exponential window, however, preserves convolution and provides taper at the segment finish. For this, and for other reasons described below, exponential weighting was chosen as the window function. In order to reduce the segmentation error at the beginning of the segment, the segments were defined such that the segment starts began only after periods of relative silence in the speech activity. This way, there was little error at the segment start from the echo of the previous speech segment, although such error could not usually be eliminated entirely. Speech pauses occur quite frequently and have durations of about 0.1 to 0.2 seconds [4]; therefore, the segmentation scheme adopted should remove segmentation error most effectively for room responses concentrated within several hundred milliseconds. The effect of the exponential windowing and the segment selection strategy was to decrease substantially the cepstral noise encountered due to segmentation errors.

Multiplication by the exponential window function $w(n) = \gamma^n$, for $|\gamma| < 1$, also has the effect of moving z-plane zeroes inward radially and hence, for sufficiently small $|\gamma|$, of converting mixed phase impulse responses to minimum phase [1]. It is easier to deal with minimum phase sequences for computational reasons, for their lack of linear phase ambiguities [1], and for their greater separability from speech in the cepstral domain [1]. Although it is desirable to use heavy weighting (small values of γ) to ensure that the impulse response becomes minimum phase, the degree of weighting which can be applied was found to be limited. A weighting value which did not introduce additional distortion at cepstral values up to several hundred milliseconds was $\gamma = 0.999$ (for 8 kHz speech corresponding to a time constant of 125 ms).

The complex cepstrum was computed from the log magnitude of the spectrum. Phase unwrapping, required for mixed phase sequences to resolve phase ambiguities, was therefore avoided. In cases where the exponential weighting used was insufficient to convert the room impulse response to minimum phase, the resulting filtered speech had a slight chime distortion similar to speech processed with filters designed to remove minimum phase magnitude equivalent impulse responses [5].

The computed cepstrum was averaged over several segments. This had the effect of reducing cepstral noise caused by remaining segmentation error and of reducing the background cepstral level due to speech. Furthermore, since the pitch of the speech was not exactly constant over the speech record, the large cepstral peak at the pitch period [1] was "smeared" and reduced by averaging. This allowed cepstral peaks due to the reverberation impulse response around the pitch period to be identified by peak-picking. Thus even for cases when the normal cepstral separation assumptions could not be made, identification of the echo cepstrum could be made. Performance was usually best when only the range corresponding to the first 15 ms (corresponding to the maximum expected pitch period) was peak-picked. We discuss this point further in the following section.

The averaged, peak-picked cepstrum was transformed to the time domain, and exponential de-weighting was applied to provide an estimate of the impulse response. From the estimated impulse response, a least squared-error filter was designed. This technique is described in [6]. The impulse response estimates were in general mixed phase and hence filters with delays were specified. Best perceptual results were achieved with filter lengths on the order of the impulse response durations and with short delays.

Because the overall filter structure contains no delays between the reverberant and processed speech other than those required in the filtering step, it was felt that the proposed filter structure could be suitable for real-time operation. In this mode, the cepstral processing would operate as a background procedure for which the processing delay, at least as large as the segment buffer size, would affect only the "up-to-dateness" of the filter coefficients.

3. RESULTS

The above procedures were tested using approximately ten seconds of speech digitized at 8 kHz and convolved with simulated room impulse responses generated with the image model [7]. In Figure 2 a simulated impulse response of an 6.4m x 6.4m x 4.2m enclosure with source-microphone distance of 0.92m is shown. The reflection coefficients of the walls are 0.9, and those of the floor and ceiling are 0.4. The impulse response, truncated at 128ms, is mixed phase and has 32 z-plane zeroes outside the unit circle and 992 zeroes inside. From the resulting reverberant speech 11 segments of duration 4096 samples were selected by examination of the reverberant speech waveform. The complex cepstrum was calculated using FFT's of length 8192 samples. The exponential weighting factor used, $\gamma = 0.999$, was in this case not sufficiently small to move all zeroes inside the unit circle. Thus it represents a "real" scenario in which the required exponential weighting factor would not be known beforehand.

Peak picking after linear cepstral scaling was performed in the cepstral region $0 < n \leq 150$. Figure 3 shows the estimated impulse response calculated from the reverberant speech. The estimate was truncated at 600 samples and was used to design an 800 tap least squares filter with delay 200 taps. The corresponding convolution of the filter with the impulse response of Figure 2 is shown in Figure 4. The large, early "spikes" were greatly reduced with the application of the filter but some new error in the form of low amplitude, long delay echo was introduced. Listening tests showed that the filtered speech had much less reverberant "boomy" sound, but tone-like distortion

was noticeable. The ratio of direct to reverberant energy for the "enhanced" impulse response was 6.0 dB vs 1.7 dB for the original.

This method was further tested with different impulse responses. The best results were achieved for impulse responses which had few z-plane zeroes far outside the unit circle, for which "light" exponential weighting was sufficient for conversion to minimum phase. We found that weighting heavily led to distortion in the estimated impulse response which increased with delay. Futhermore, better enhancement was achieved for impulse response which were characterised by large, discrete peaks. In these cases, cepstral peak-picking could be extended throughout the entire cepstral range and less distortion was evident in the estimated impulse response.

3.1 Discussion of Results

This dereverberation technique presupposes that the impulse response is made minimum phase by exponential weighting. However, choosing a small value of γ designed to accommodate all expected impulse responses led to distortion upon exponential de-weighting. We speculate that as γ is reduced, the time window falls off more sharply and the beginning of the segment is emphasized. Therefore any residual segmentation error which is not corrected by choice of segment start location will become magnified for smaller values of γ.

For lighter weighting using values of γ closer to unity, experimentation revealed that the segmentation distortion had been largely removed. However, residual speech cepstrum remaining at all cepstral locations after averaging represented a limit to the performance of this technique. The speech cepstrum was reduced, but not eliminated, by increasing the number of segments averaged. In cases where the reverberation impulse response cepstrum was large compared with the residual speech cepstrum, the enhancement procedure was effective. If the echo cepstrum was of a form suitable for peak-picking, the enhancement was also effective and represented a large improvement over the techniques described in [2]. In general, however, the room impulse responses which we generated to represent typical office scenarios with the program of [7] were not of a form suitable to peak-picking.

4. CONCLUSIONS

The approach to speech dereverberation taken in this paper was to estimate the reverberation impulse response via cepstral techniques and to remove the estimated impulse response with a linear inverse filter. During the estimation step, speech pauses are exploited to reduce segmentation error and cepstral averaging is used to reduce the background speech cepstrum. However, residual segmentation error prevents the use of heavy exponential weighting required for some room impulse responses. Possible solutions to this problem would involve a two-step or alternately a closed-loop approach in an attempt to further reduce segmentation error. The presence of background speech cepstrum after averaging limits the estimation accuracy. At present, therefore, this cepstral method is not suited for application to room impulse responses which have zeroes far outside the unit circle or which have dispersed, low-level cepstra.

References

1. A. Oppenheim and R. Schafer, *Digital Signal Processing*, Prentice-Hall, New Jersey 1975
2. R. Schafer, "Echo Removal by Discrete Generalized Linear Filtering", *MIT Technical Report #466*, 1969
3. J. Tribolet, "A New Phase Unwrapping Algorithm", *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-25, No. 2 1977
4. M. Picheny, N. Durlach, and L. Braida, "Speaking Clearly For the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech", *Journal of Speech and Hearing Research*, Vol 29, Dec 1986
5. S. Neely, J. Allen, "Invertibility of a Room Impulse Response", *Journal of the Acoustic Society of America*, Vol. 66, No. 1 1979
6. J. Mourjopoulos, "On the Variation and Invertibility of Room Impulse Response Functions", *Journal of Sound and Vibration*, Vol. 102, No. 2 1985
7. J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *Journal of Acoust. Soc. Am.*, Vol. 65 No. 4 1979

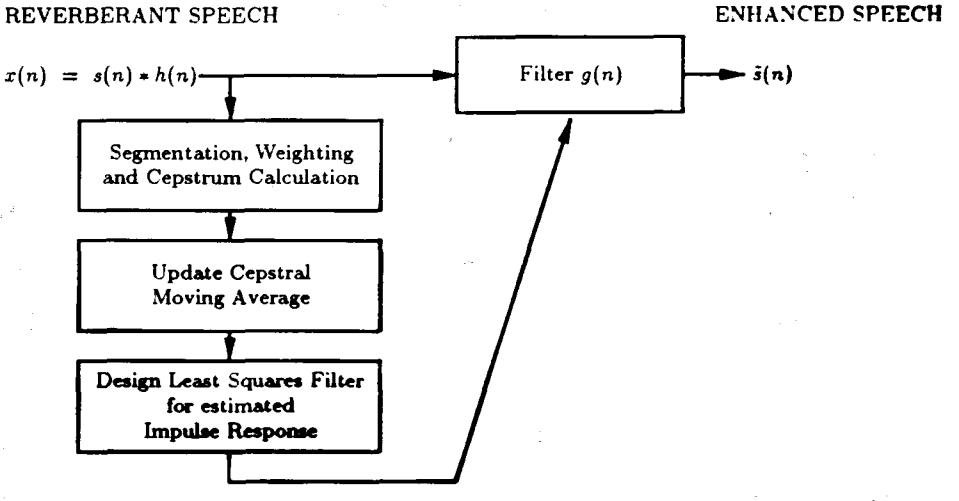


Fig. 1 Dereverberation System Block Diagram

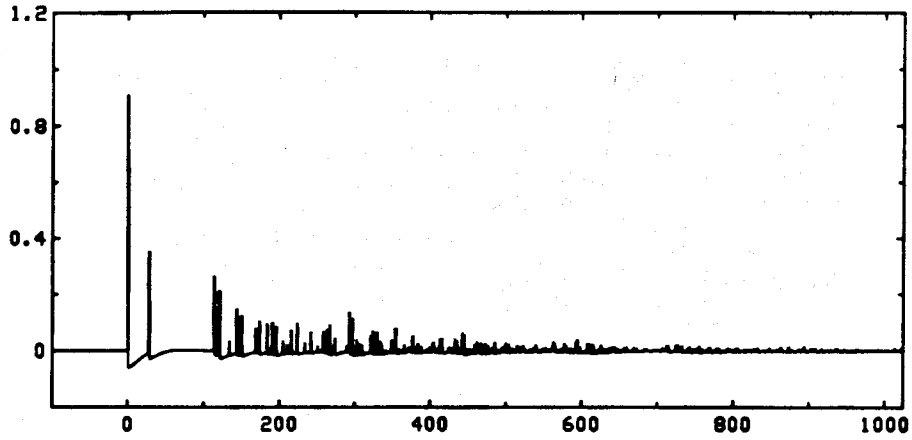


Fig. 2 Simulated impulse response

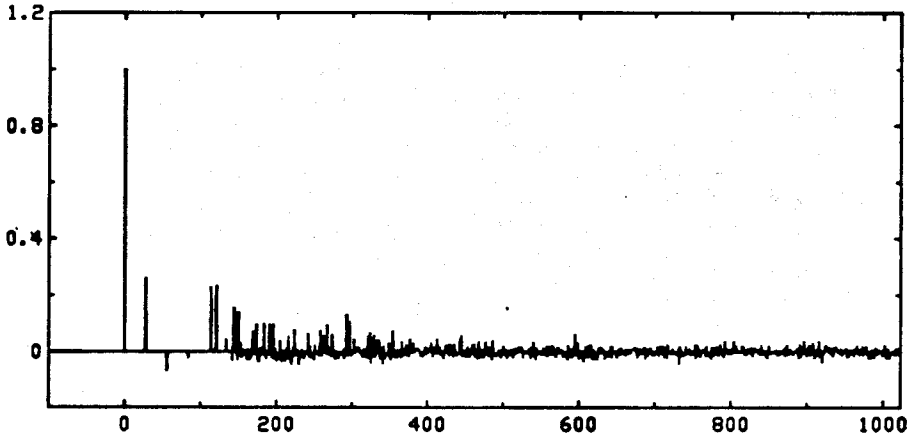


Fig. 3 Estimated impulse response

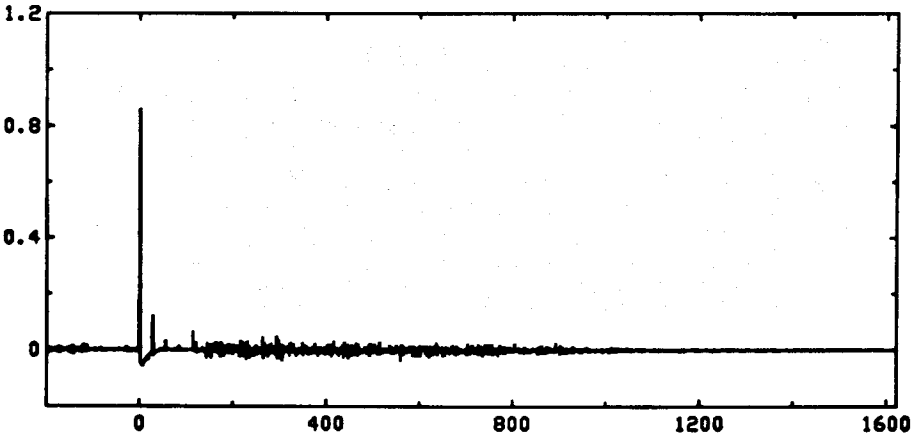


Fig. 4 Convolution of filter and simulated impulse response