S14.25

REVERBERANT SPEECH ENHANCEMENT USING CEPSTRAL PROCESSING

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

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

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

ABSTRACT

Complex cepstral deconvolution is applied to acoustic dereverberation. It is found that traditional cepstral techniques fail in acoustic dereverberation because segmentation errors in the time domain prevent accurate cepstral computation. An algorithm for speech dereverberation is presented which incorporates a new approach to the segmentation and windowing procedure for speech. Averaging in the cepstrum is exploited to increase the separation between speech and impulse response. An estimate of the room impulse response is built up, and a least squared error inverse filter is used to remove the estimated impulse response from the reverberant speech. Reduction of reverberation with this technique is demonstrated.

1. INTRODUCTION

The dereverberation of acoustically reverberant speech has potential application to the enhancement of speech which has become degraded through the addition of multiple echoes. For example, the "hands-free" telephone, which is used widely in office rooms, often suffers from reverberation when the microphone is placed too far from the talker. In this case, the ratio of echoed speech reflected from walls and other hard surfaces to direct path speech becomes large, and the far-end listener perceives the speech as reverberant. Mathematically,

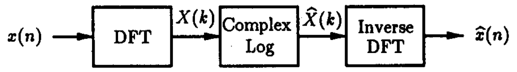
$$x(n) \;=\; s(n) * h(n) \qquad (1)$$

where $s(n)$ is sampled time signal representing the "clean" speech, $h(n)$ is the impulse response of the room, and $x(n)$ is the sum of the direct path speech and the resulting echoes.

Most successful techniques for processing reverberant speech have relied upon measuring two (or more) signals $x_1(n)$ and $x_2(n)$ at different room locations and exploiting the uncorrelatedness between $h_1(n)$ and $h_2(n)$. Single microphone reverberant speech enhancement typically requires prior knowledge of $h(n)$ and subsequent inverse filtering. The impulse response may be measured from the response to a known signal; techniques for estimation of $h(n)$ from the reverberant speech alone have not been described. Another possible single microphone approach is the application of complex cepstral filtering to $x(n)$. In this approach, utilized in [1] to process speech degraded by simple echoes (see also [2]), no knowledge of $h(n)$ beyond certain broad assumptions is required to deconvolve it from $s(n)$. In this paper, we present the results of a study into the use of cepstral techniques for enhancement of continuous speech which has been subject to simulated room-reverberation. We have found that a direct application of the techniques described in [1] is unsuitable for reverberant speech enhancement. However, by modifying the windowing procedure, and by using cepstral averaging to identify $h(n)$ before subsequent inverse filtering, we are able to achieve significant reduction in reverberant effect.

1.1 Cepstral Filtering Techniques

In this section we provide a brief review of the complex cepstrum and the techniques by which a segment of speech may be separated from a convolved impulse response representing a simple pattern of echoes. The complex cepstrum is described in [3]. 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 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 an impulse response which in the log spectrum is characterized by rapid "ripples", and which in the complex cepstrum is composed of pulses concentrated far away from the cepstral origin.

Schafer [1] developed procedures whereby the complex cepstrum is calculated from segments of reverberant speech to which an exponentially weighted window function is applied, 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, and multiplied by the inverse of the exponential window to form the enhanced speech.

We found, however, that when these techniques were applied to the dereverberation of speech subject to acoustic reverberation, the complex cepstral method led to distortion or incomplete dereverberation in the processed speech. Upon investigation, it was revealed that the most serious problems were related to the process of segmentation of $x(n)$. In excising a finite length segment $x_i(n)$ from the signal of indefinite extent $x(n)$, one obtains a signal that can only approximately be represented by the convolution of $h(n)$ with some clean-speech segment $s_i(n)$. Thus, the complex cepstrum $\hat{x}_i(n)$ can not be said to be the addition of $\hat{h}(n)$ and $\hat{s}_i(n)$. Following this line of reasoning, any cepstral windowing or peak removal operations on $\hat{x}_i(n)$ can not be expected to remove $h(n)$ completely, and can be expected to distort the resulting estimate $\tilde{s}_i(n)$. In [1], a method is presented whereby the estimated segments $\tilde{s}_i(n)$ can be joined in such a way as to account for the effect

[†] Now with Bell Northern Research in Montreal, Canada

CH2977-7/91/0000-0977 $1.00 © 1991 IEEE

of segment truncation error. However, this method assumes that $\widehat{h}(n)$ has been calculated correctly, and that segmentation error need only be considered for the segment reconstruction process. Our key finding for acoustical dereverberation is that this assumption is not appropriate. In the next sections, we consider the windowing and segmentation problem in more detail.

2. WINDOWING AND SEGMENTATION

In this section, we investigate the effect of the segmentation procedure and of the time-domain window function upon the calculation of the complex cepstrum. Let us begin with the choice of window function. We considered rectangular, Hamming-type, and exponential window functions as candidates. Each segment of reverberant speech may be written [1] as

$$x_i(n) = s_i(n) * h(n) + e_i(n) \tag{2}$$

We may describe the error term as the sum

$$e_i(n) = v_i(n) - u_i(n) \tag{3}$$

where $v_i(n)$ is the "extra" echo which intrudes from the previous segment and $u_i(n)$ is the "missing" tail of the echo of the speech of the current segment. Intuitively, the goal of windowing would be to reduce the importance of these error components by smoothly tapering the segment boundary, while at the same time not introducing distortion into the calculated cepstrum.

Rectangular windows obviously provide no tapering at segment boundaries. Functions such as Hamming windows are tailor-made for reduction of truncation error, but their effects upon the convolutional combination of signals (of extent on the same order as the window) are not known. Consider the windowed signal, $y(n) = x(n)w(n)$ for which the spectrum is

$$\begin{aligned} Y(\omega) &= [S(\omega)H(\omega)] * W(\omega) \\ &= \int_{-\pi}^{\pi} [S(\lambda)H(\lambda)]W(\omega-\lambda)d\lambda \end{aligned} \tag{4}$$

For the Hamming window $w(n) = 0.5[0.54 - 0.46\cos(\frac{2\pi n}{N-1})]$,

$$Y(\omega) = 0.5\left[.54X(\omega) - .23X(\omega - \frac{2\pi}{N-1}) - .23X(\omega + \frac{2\pi}{N-1})\right] \tag{5}$$

It is extremely difficult to find an expression for the complex logarithm of equation (5) which will allow us to predict the effect upon the cepstrum.

Exponential windows, with $w(n) = \gamma^n$, where $|\gamma| < 1$, provide taper at the segment finish only. However, because they do not destroy the convolutional combination between signals, they affect the cepstrum in a known way. As shown in [3], for $x(n)$ as in (1),

$$\gamma^n x(n) = \gamma^n s(n) * \gamma^n h(n) \tag{6}$$

Thus the complex cepstrum of the exponentially weighted signal remains a sum of the cepstra of two convolutionally combined components.

To check which of the above windows would be most suitable for calculation of the cepstrum in the presence of segmentation error $e_i(n)$, we constructed sequences of white noise as the "speech" signal $s(n)$. These were convolved with an impulse response representing a simple, minimum phase echo of amplitude 0.5 and delay 200 samples. This echo has an entirely causal complex cepstrum consisting of a series of pulses at $n = 200, 400, \ldots$ with the amplitude at the first pulse $\widehat{h}(200) = 0.5$ [3]. (With exponential weighting of 0.996^n, the pulse amplitude is changed to $(0.996^{200}) \times 0.5 = 0.224$ [3]). From the echoed signal were then excised segments of various lengths N, analogous to truncated speech segments $x_i(n)$, and one of the three window functions was applied. The value of the resulting complex cepstrum $\widehat{y}_i(n)$ at the two points $n = \pm 200$, averaged over a small number of trials, is reported in Table 1. First note that the cepstral contribution of the signal $s_i(n)$ was negligible at these locations. Three immediate observations can then be made from Table 1. First, it can be seen that for rectangular and Hamming windows, spurious pulses are encountered at $n = -200$. Second, for all window functions, the value of cepstrum at $n = 200$ is substantially lower than $\widehat{h}(200)$. Further experiments revealed that with non-minimum phase echoes, the cepstrum was equally distorted. Also, for maximum phase echoes, the exponential window produced a pulse value at $n = 200$ much smaller than would be predicted from theory. Third, the accuracy of the computation does not improve with increasing window length. Similar results were also observed with actual speech samples rather than white noise.

WINDOW	n	$N = 512$	$N = 1024$	$N = 2048$	$N = 4096$
Rectang	−200	0.123	0.183	0.156	0.211
	+200	0.120	0.174	0.256	0.240
Hamming	−200	0.078	0.174	0.185	0.231
	+200	0.077	0.175	0.263	0.267
Expon	−200	0.006	−0.002	0.008	0.002
$\alpha = 0.996$	+200	0.152	0.185	0.185	0.173

Table 1 Average cepstral peaks calculated for various window types and lengths

We speculate that both Hamming windowing and non-tapered segmentation errors of the rectangular window cause the phase curve of the complex logarithm to be distorted, resulting in confusion between minimum and maximum phase components of the impulse response cepstrum. A similar effect was noted in [4]. Therefore, we cannot use either of these windows for acoustic dereverberation since we expect, in general, a non-minimum phase room impulse response and segmentation errors in $x_i(n)$.

As for the exponential window, we speculate that segmentation error $v_i(n)$ appearing at the segment start causes cepstral distortion. Although the mechanism remains unclear, it also appears that confusion between maximum phase and minimum phase components persists with exponential windows in the presence of segmentation error in the following sense: when the same experiment was run using a maximum phase echo of amplitude 2.0 (which corresponds to the maximum phase impulse response with the same spectral magnitude as the minimum phase echo of amplitude 0.5), a very similar cepstral value at $n = 200$ to that in the minimum phase case was computed. Under the given exponential weighting, the value of the cepstral pulse should be $0.996^{200} \times 2.0 = 0.897$.

Thus, exponential windows are also of little use in the presence of segmentation error. In order to ameliorate the situation, we proposed the following: define a segment to begin only after a period of speech silence, and apply an exponential window. In this way, $s_{i-1}(n) \approx 0$. Speech pauses occur quite frequently and have durations of about 0.1 to 0.2 seconds [5]; therefore, the segmentation scheme adopted should remove most segmentation error $v_i(n)$ for room responses concentrated within several hundred milliseconds. We supposed that the smooth taper to zero of the exponential window would "remove" segment end error $u_i(n)$, for large enough values of N. Preliminary tests revealed that this strategy was very effective. In the next section, we outline a dereverberation algorithm based upon this windowing and segmentation technique.

3. DEREVERBERATION ALGORITHM

In the proposed dereverberation algorithm, the primary goal is to estimate accurately the reverberation impulse response. From this estimate, a least squares filter is designed and applied to the reverberant speech. Figure 1 shows a block diagram of the dereverberation system, which is structured so that it may be run in real time. In the first step, the reverberant speech is segmented according to the procedure described above. For the purposes of this research, this segmentation was

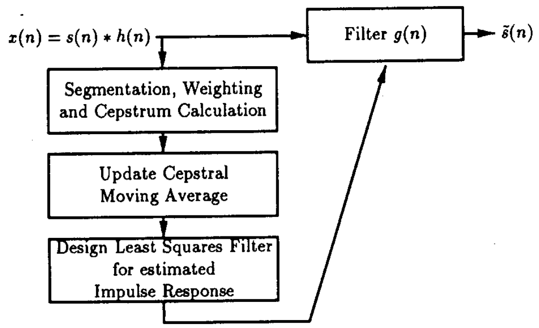

$x(n) = s(n) * h(n)$
Filter $g(n)$
$\tilde{s}(n)$
Segmentation, Weighting and Cepstrum Calculation
Update Cepstral Moving Average
Design Least Squares Filter for estimated Impulse Response

Fig. 1 Dereverberation System Block Diagram

done "by eye" on the reverberant speech waveforms, but automation should be straightforward.

Next, exponential weighting of each segment is performed. Note that in addition to the tapering effect of the exponential window, multiplication by $w(n) = \gamma^n$, 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 [3]. It is easier to deal with minimum phase sequences because of their lack of linear phase ambiguities and because of their greater separability from speech in the cepstral domain. Also, for minimum phase sequences, phase unwrapping is not required in the computation of the complex logarithm. Accordingly, the complex cepstrum is then computed from the log magnitude of the spectrum. It is necessary to make the assumption that the room impulse response is converted to minimum phase with the exponential weighting factor chosen.

The computed cepstrum is averaged over several segments. This has 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 is not exactly constant over most speech records, the large cepstral peak at the pitch period [3] becomes "smeared" and reduced by averaging. An example of the beneficial effects of cepstral averaging is shown in Figure 2, where the pitch peak can be seen to wander while the cepstrum due to $h(n)$ remains constant. This allows cepstral components due to the impulse response located around the pitch period to be identified by peak-picking. Thus even for cases when the normal cepstral separation assumptions (that impulse response components are not located around the pitch period) cannot be made, identification of the echo cepstrum can proceed. We found that performance was usually best when only the range within the first 15 ms (corresponding to the maximum expected pitch period) was peak-picked.

The averaged, peak-picked cepstrum is transformed to the time domain, and exponential de-weighting is applied to provide an estimate of the impulse response, which is truncated at an appropriate length. From the estimated impulse response, a least squared-error filter is designed. This technique is described in [6]. The impulse response estimates are in general mixed phase and hence filters with delays are specified. From Figure 1 it can be seen that the only delay between the reverberant and the processed speech is this filtering delay, making the algorithm potentially suitable for real-time operation. The delays involved in the cepstral processing affect only the "up-to-dateness" of the filter coefficients. Best perceptual results were achieved with filter lengths on the order of the impulse response durations and with short delays, on the order of $\frac{1}{4}$ of the filter length.

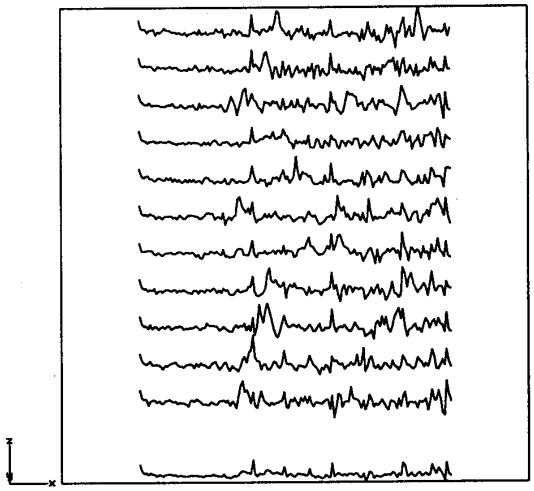


Fig. 2 Average over 11 speech segments of first 150 cepstral values; the trace at bottom is the average of the 11 upper traces

4. 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 3 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 4 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 3 is shown in Figure 5. 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 low level, 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. For impulse responses formed of a number of discrete, well defined peaks, the processed speech was remarkably superior to the reverberant speech. Such impulse responses are not, however, characteristic of those encountered in typical rooms.

In general, the best results using simulated room impulse responses were achieved for 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 echo delay.

4.1 DISCUSSION

We have presented some results of a study into the applicability of cepstral processing to reverberant speech enhancement. We have found that direct deconvolution using the methods of [1] is not practical. The accuracy of the computed cepstrum is critically dependent upon the degree of segmentation error in each reverberant speech segment. The accuracy of the cepstral computation is equally degraded when tapered windows such as Hamming functions are applied. We therefore use exponential weighting to introduce taper at the segment end, and reduce segment start error by choosing segment starts to begin after silent periods. Cepstral averaging then allows accurate identification of the reverberation impulse response.

This 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 leads 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 becomes magnified for smaller values of γ. Possible solutions to this problem would involve a two-step or alternately a closed-loop approach which would remove some reverberation before applying a heavier exponential weighting. In this way, segmentation error would be decreased and would lead to less distortion upon heavy exponential weighting.

For weighting using values of γ closer to unity, experimentation revealed that the segmentation distortion is largely removed. However, residual speech cepstrum remaining at all cepstral locations after averaging represents a limit to the performance of this technique. In cases where the reverberation impulse response cepstrum is large compared with the residual speech cepstrum, the enhancement procedure is effective. If the echo cepstrum is of a form suitable for peak-picking, the enhancement is also effective and represents a large improvement over the techniques described in [1]. The residual speech cepstrum may be reduced by increasing the number of segments averaged. This, of course, implies that the impulse response must remain constant over the averaging period. Further research is required to determine the optimal tradeoff between these factors.

References

1. A. Oppenheim, R. Schafer, T. Stockham, "Nonlinear filtering of multiplied and convolved signals", *IEEE Trans. Audio and Electroacoustics*, Vol. AU-16, No. 3, 1968.
2. R. Schafer, "Echo removal by discrete generalized linear filtering", *MIT Technical Report #466*, 1969.
3. A. Oppenheim and R. Schafer, *Digital Signal Processing*, Prentice-Hall, 1975.
4. J. Tribolet, T. Quatieri, A. Oppenheim, "Short-time homomorphic analysis", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Hartford, 1977.
5. 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, No. 4, 1986.
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.

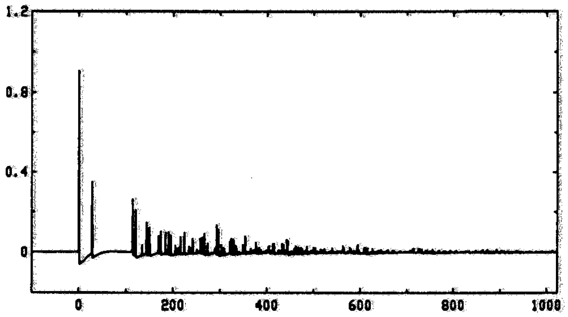


Fig. 3 Simulated impulse response

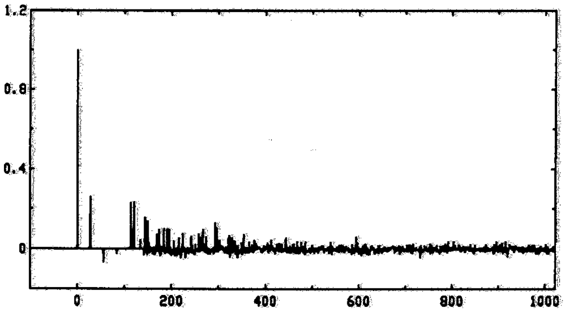


Fig. 4 Estimated impulse response

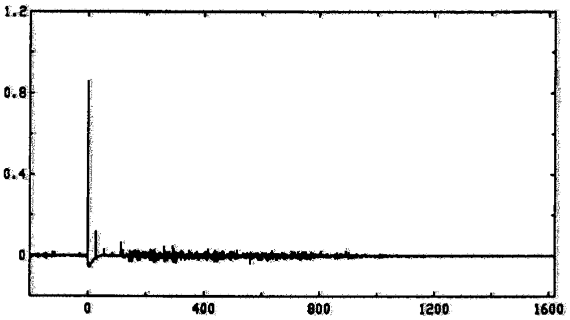


Fig. 5 Convolution of filter and simulated impulse response