ALTERNATIVE METHODS FOR REVERBERANT SPEECH ENHANCEMENT

D. Cole, M. Moody and S. Sridharan

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT — A novel method for enhancement of reverberant speech is presented. The technique, which is based on the well known spectral subtraction enhancement method, overcomes the positional sensitivity which makes the conventional estimate/invert method impractical.

INTRODUCTION

The conventional method for enhancement of single channel reverberant speech is a two step procedure: first estimating the room response $h(n)$; then designing and applying the inverse filter $h'(n)$. Of the two steps, the latter is more straightforward, and may be performed using either least-squared error or complex cepstral separation techniques (Mourjopoulos *et al.*, 1982). This paper will demonstrate the susceptibility of this method to positional change by using a small variation in speaker position and orientation — this produces very large performance degradation.

The inability of the conventional approach to deal with positional variation is quite frustrating, as there is not a great audible change between positions. We suggest that alternative techniques might exploit the relative positional invariance of spectrally equivalent minimum phase response, and describe a spectral subtraction based procedure which is not degraded by positional variation.

Motivation for this research is its usefulness in forensic applications, where use of a single microphone is a common constraint. Choice of microphone placement may be very limited in such circumstances, and severe reverberation characteristics often cause great difficulty in understanding the recorded speech, even with low background noise levels. Obviously, speaker position cannot be constrained in such situations, so a reverberant speech enhancement method which is robust to positional changes would be of great value.

Acoustical reverberation of speech may be described mathematically as the convolution of the 'clean' speech signal $s(n)$ with the impulse response of the room $h(n)$:

$$x(n) = s(n) * h(n) \tag{1}$$

Single microphone enhancement of reverberant speech typically requires estimation of $h(n)$ to derive an inverse filter (Mourjopoulos *et al* 1982). The greatest difficulty in this approach is in obtaining an accurate estimate where the response to a known signal cannot be obtained. This is the usual case for covert recordings. The difficulty is compounded by the large variations of $h(n)$ with positional changes of either source or receiver (Mourjopoulos 1985). Classical techniques of homomorphic deconvolution generally perform poorly for reverberant speech for several reasons: framing effects, cepstral overlap of speech and echo, generally mixed phase signal characteristics, and phase unwrapping difficulties when the complex cepstrum is used for reconstruction.

Bees, Blostein and Kabal (1991) described a modified cepstral approach for estimation of $h(n)$ from the reverberant speech signal only, by exponentially windowing speech segments to force a minimum phase characteristic and averaging of a number of such frames. This appears to be the most successful estimation approach to date, but its practical use was shown by Cole *et al.*(1994) to be limited severely by the requirement for a minimum phase characteristic after windowing and by variation of $h(n)$ with positional changes.

ROOM CHARACTERISTICS

The impulse response of a bare room (of about $40m^3$, with reverberation time of about 3 seconds) was measured using a chirp signal for various speaker positions and orientations, as outlined in Cole *et al.*(1994). A loudspeaker and microphone were placed in the room used, about $4m$ apart. The room impulse response was found by using a chirp signal for three slightly different speaker orientations: a reference position, with the speaker rotated 45° and with the speaker shifted laterally by $0.5m$. These impulse responses were then used for all testing. (Note that this avoids the need for blind estimation of the room response, which is not feasible for this length of response using the usual techniques.) The reference response was used to artificially reverberate clean speech utterances for articulation testing.

ROOM RESPONSE INVERSION

The reference and rotated position responses were used to evaluate inversion of room response for two cases:

1 when a perfect estimate of the room response is available; and

2 when the source moves suddenly, as might be typical in a real situation.

An inverse filter was designed for the reference position using a least mean squares error criterion. The first case was then evaluated by using the resultant response of the convolution of the reference impulse response and the reference inverse filter. The second case was evaluated by using the resultant of the convolution of the rotated position impulse response and the reference inverse filter. Two evaluation methods were used:

- The U_{80} useful-detrimental ratio as evaluated by Bradley (1986) was calculated as the ratio of energy arriving in the first 80ms to the energy arriving after. Bradley found this to be as good a predictor as more complex ratios in his study of auditoria.
- An articulation test, using sets of phonetically balanced (PB) words embedded in a carrier phrase, was conducted with twelve listeners, each a native Australian English speaker with no known hearing impediment. The choice of a PB word test follows the findings of Kruger, Gough and Hill (1991), who found this the most suitable method for testing intelligibility of reverberant speech. The PB word lists used were those of Clark (1981), for Australian English.

These results are summarised in Table 1, which indicates extremely good performance when the room response estimate is accurate, but extreme degradation due to movement. In this case, a 45° rotation (which might be typical of a turn of the head) degrades the result to a point where it is less intelligible than the unprocessed reverberant speech. This result suggests that the inversion technique can only be feasible if the room response estimating procedure can accurately track changes of source position. Such an estimating procedure does not now exist, and the task seems intractible.

The results are further illustrated by Figure 1, which shows an example of waveforms of the clean and reverberant speech, and the processed speech for the two positions. Subjectively, the reference position result sounds identical to the clean speech, with a low level of additive noise, while the presence of speech is audibly discernible in the rotated position result, but obscured by a high level of noise apparently uncorrelated to the speech.

	AS (% correct)	U_{80} (dB)
(a) Raw	29	-4.4
(b) Reference	98	16.7
(c) Rotated	21	-10.1

Table 1: Articulation score AS (% correct) and useful - detrimental ratio U_{80} (dB) *for (a) raw reverberant speech, (b) inverse filtered speech in the reference position and (c) inverse filtered speech in the rotated position.*

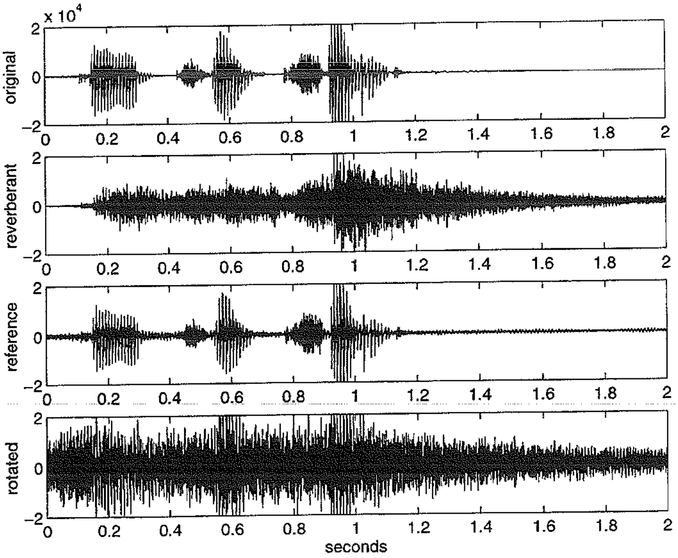

Figure 1: Room response inversion waveforms
Original speech; reverberant speech; resultant for reference position; resultant for rotated position.

ALTERNATIVE APPROACHES

The inability of the conventional inversion approach to deal with positional variation is quite frustrating, as there is not a great audible change between positions. The change of impulse response with position is mostly due to perturbations in relative times of arrival of reflections, producing phase variations which are relatively insignificant to the human auditory system, but are catastrophic to the mathematical deconvolution procedure used for inverse filtering.

The spectral magnitude characteristics of room responses have been reported to be relatively constant (Cole *et al.*1996). This is shown by examination of the smoothed spectrum of short portions of the three impulse responses at similar temporal displacements. Figure 2 shows the LPC smoothed spectrum (offset by 10dB for clarity) at several points in time along the impulse responses.

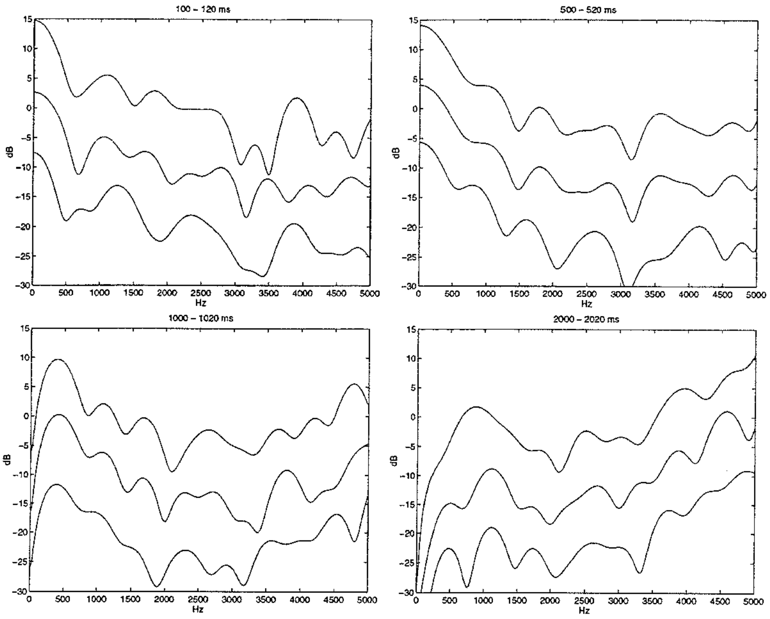

Figure 2: Smoothed spectral magnitudes, measured room responses
upper: reference position
centre: rotated position
lower: shifted position

The similarity of the smoothed spectra of the impulse responses corresponds with the subjective similarity of reverberation produced in the different positions. It is possible that this similarity might be exploited in a more robust enhancement method. We would expect that the energy spectrum of the reverberant component of the signal will be quite similar over a range of positions. This suggests a spectral subtraction type approach, where estimates of the spectral magnitude of the reverberant signal component are subtracted from that of the signal at the appropriate time offsets. While such an approach would not produce as much improvement as the ideal estimation–inversion approach, it should be more robust to positional change.

Other approaches might also be found to exploit the relatively low effect of position on spectral magnitude characteristics of the reverberant signal. For example, we speculate that the linear prediction coefficients of the reverberant signal should not vary greatly with position. A method of estimating the LPC vectors of the clean speech from those of the reverberant speech (effectively reconstructing the correct vocal tract response) should thus be quite robust with respect to positional change.

A MODIFIED SPECTRAL SUBTRACTION APPROACH

To evaluate the feasibility of the approaches proposed, a modified spectral subtraction technique was devised. The reference position impulse response was assumed known, and was segmented into N frames of about 20msec (corresponding to the well accepted frame size for pseudo-stationarity of speech.) Each of these frames was analysed to obtain the frame energy content E_n and linear

predictive coefficient vector A_n.
It was assumed that the normalised frame energy content E_n/E_0 and the LPC vectors A_n are independent of source position. These parameters were then used to estimate the reverberant signal component due to each impulse response frame and a spectral subtraction procedure used to remove it. This further assumes that speech segments are independent and uncorrelated — an assumption which obviously is often poor.

This is an iterative procedure, starting from the 'tail' of the impulse response. The initial signal estimate $\hat{x}_0$ is set to the reverberant signal. Then (with k a small constant),
for $n = 1$ to $N - k$:

$$\hat{x}_n = \mathsf{SS}_n \left(\hat{x}_{n-1}, \left(\frac{E_{N-n}}{E_0} \right)^{\frac{1}{2}} (\hat{x}_{n-1} * A_{N-n}) \right)$$

where $\mathsf{SS}_m(p, q)$ denotes the spectral subtraction from signal p of signal q offset by m frames.

The U_{80} useful-detrimental ratio has been calculated for each of the three positions as shown in Table 2. No articulation scores have been taken yet for the modified spectral subtraction method. Figure 3 shows an example of waveforms of the clean and reverberant speech, and the processed speech for two of the positions (corresponding to those of Figure 1.) Subjectively, the results for all positions sound very similar. Although they are not of the quality of the reference inversion result, the procedure is highly robust to the positional changes tested.

	U_{80} (dB)
(a) Raw	-4.4
(b) Reference	10.2
(c) Rotated	12.1
(d) Shifted	10.9

Table 2: Useful - detrimental ratio U_{80} (dB)
for raw reverberant speech and modified spectrally subtracted speech in the reference, rotated and shifted positions.

CONCLUSIONS

The results presented show the shortcomings of the conventional estimate - invert paradigm for enhancement of reverberant speech. Blind estimation procedures available cannot adequately estimate room responses in a static situation. Where either source or receiver is moving, the problem is magnified by the extreme phase sensitivity of the response.

The iterative spectral subtraction procedure detailed provides a method which is robust to positional change. The method used suffers from the very crude spectral subtraction method used, but this is an aspect which might be improved in numerous ways.

The key advantage of the new technique is that it requires only an approximate impulse response which can be found by measurement at a later time, and it does not require phase information. By contrast, the conventional inversion approach requires a very accurate impulse response estimate, including phase information, at all times.

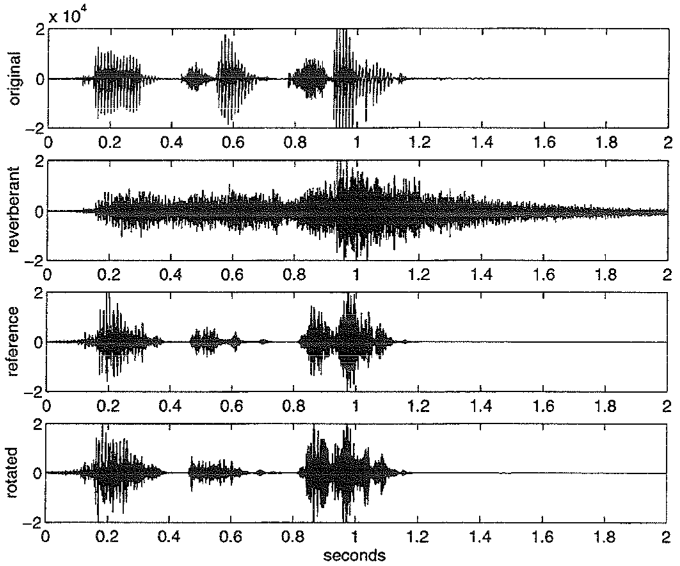

Figure 3: Modified spectral subtraction waveforms
Original speech; reverberant speech; resultant for reference position; resultant for rotated position.

REFERENCES

Bees D., Blostein M., Kabal P. (1991) "Reverberant Speech Enhancement using Cepstral Processing", IEEE ICASSP-91, 977–80.

Bradley J.S. (1986) "Predictors of speech intelligibility", J.Acoust. Soc. Am. **80** (3), 837–45.

Clark J.E. (1981) "Four PB word lists for Australian English", Australian Journal of Audiology **3** (1), 21-31.

Cole D., Moody M., Sridharan S. (1994) "Intelligibility of reverberant speech enhanced by inversion of room response", IEEE ISSIPNN'94, 241–4.

Cole D., Moody M., Sridharan S. (1996) "Enhancement methods for reverberant speech", IEEE Int. Symp. Sig. Proc. App., 383–6.

Kruger K., Gough K., Hill P. (1991) "A comparison of subjective speech intelligibility tests in reverberant environments", Canadian Acoustics **19** (4), 23-4.

Mourjopoulos J. (1985) "On the variation and invertibility of room impulse response functions", Journal of Sound and Vibration **102** (2), 217–228.

Mourjopoulos J., Clarkson P., Hammond J. (1982) "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals", IEEE ICASSP-82, 1858-61.