ROBUST ENHANCEMENT OF REVERBERANT SPEECH USING ITERATIVE NOISE REMOVAL

David Cole (d.cole@qut.edu.au)
Miles Moody (m.moody@qut.edu.au)
Sridha Sridharan (s.sridharan@qut.edu.au)

Speech Research Lab, Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434 Brisbane, Australia

ABSTRACT

We suggest a new technique for the enhancement of single channel reverberant speech. Previous methods have used either waveform deconvolution or modulation envelope deconvolution. Waveform deconvolution requires calculation of an inverse room response, and is impractical due to variation with source or receiver movement. Modulation envelope deconvolution has been claimed to be position independent, but our research indicates that envelope restoration in fact degrades intelligibility of the speech. Our method uses the observation that the smoothed segmental spectral magnitude of the room response is less variable with position. This is used to estimate the reverberant component of the signal, which is removed iteratively using conventional noise reduction algorithms. The enhanced output is not perceptibly affected by positional changes.

1 INTRODUCTION

Enhancement of single channel reverberant speech has been addressed occasionally, usually with applications such as hands-free telephony in mind. This paper describes research aimed at forensic applications such as enhancement of covert recordings, where intelligibility improvement is the primary goal, rather than improved quality.

In forensic situations, it is quite common to encounter very severe reverberation. To simulate this, the characteristics of a room (of about 40m^3 with reverberation time of about 3 seconds) were measured for three positions, to simulate speaker movement: a reference position, a rotated position with the speaker rotated 45° and a shifted position with the speaker moved laterally about 0.5m. These impulse responses were used to reverberate test material comprising phonetically balanced word lists, as detailed in [1]. Intelligibility scores were determined using this test material with twelve listeners for the enhancement methods discussed in the following sections.

2 WAVEFORM DECONVOLUTION

Waveform deconvolution is the approach usually suggested for enhancement of reverberant speech, which may be expressed as the convolution of clean speech $s(n)$ with the impulse response of the room, $h(n)$.

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

This requires an estimate of $h(n)$, so that an inverse characteristic $\tilde{h}'(n)$ can be designed. Given a good estimate of the room response, it is quite a simple matter to find an inverse response using least-squared error methods [3]. This produces a very good recovery of the original speech.

$$\tilde{s}(n) = x(n) * \tilde{h}'(n) \tag{2}$$

The greatest difficulty in this approach is in obtaining an accurate estimate where the response to a known signal cannot be obtained. This is compounded by variation of the room impulse response with positional change of either source or receiver. Although this variation is not perceptible audibly, the effect on the mathematical inversion procedure is severe. An inverse characteristic good for a particular source and receiver position will be invalid for even very small positional changes, and will actually degrade the reverberant speech further.

This is shown in Table 1, which shows the result of inversion for two cases. For this test, an inverse filter was designed from the reference position impulse response. When applied to speech reverberated in the reference position, very good enhancement results. However, when the reference position inverse filter is applied to speech reverberated in the rotated position, the output is even less intelligible than the unprocessed reverberant speech.

Examples of the inversion technique are available in proceedings CDROM sound files for the reverberant speech [A0595S01.wav], the reference position inverted speech [A0595S02.wav] and rotated position reverberant speech inverted by the reference inverse filter [A0595S03.wav].

Condition	AS (% correct)
(a) Raw	48
(b) Reference	98
(c) Rotated	35

Table 1: Articulation score *AS* (% correct) showing effect of movement for:
(a) raw reverberant speech, (b) inverse filtered speech in the reference position and (c) inverse filtered speech in the rotated position.

3 MODULATION ENVELOPE DECONVOLUTION

The use of an envelope convolutional model was suggested by Mourjopoulos *et al* [4] and further developed in [5]. This approach represents the signal by the product of a positive envelope function $A(n)$ and a cosine modulated instantaneous phase term $\phi(n)$:

$$\begin{aligned} x(n) &= A_x(n)\cos\phi_x(n) \\ &= A_s(n)\cos\phi_s(n) * A_h(n)\cos\phi_h(n) \end{aligned} \quad (3)$$

Making the assumptions that the phase terms change sufficiently slowly with time and that the bandwidth of the envelope function is limited, Mourjopoulos *et al* proposed that the reverberant envelopes can be approximated from the convolution of the anechoic speech envelopes and the room response envelopes:

$$A_x(n) \approx 0.5\,(A_s(n) * A_h(n)) \quad (4)$$

The approach used in [4] was to design an inverse operator ${A_h}^{-1}(n)$ for the room response envelopes using exponential approximations of the $A_h(n)$. This was refined in [5] to use a least squared error method to estimate the anechoic signal envelopes $A_s(n)$. The enhanced speech was reconstructed using the original reverberant phase.

An implicit assumption in this method is that restoration of modulation envelopes is beneficial. While this is intuitive, it has not been established in the same manner as, say, the importance of short-term spectral magnitude characteristics. To test this assumption, the modulation envelope of reverberant speech was corrected by using *a priori* information from the original unreverberated speech. This corresponds to ideal restoration of modulation envelopes. The intelligibility results shown in Table 2 indicate that modulation envelope correction in fact degrades intelligibility, even though reverberation is subjectively reduced by the technique. The reduction of intelligibility might be due to the presence of reverberant artifacts without surrounding context.

The effect of ideal modulation envelope restoration is presented in proceedings CDROM sound files containing reverberant speech [A0595S04.wav] and envelope corrected reverberant speech [A0595S05.wav] for the utterance "soap bang dot".

Condition	AS (% correct)
(a) Raw	56
(b) Amplitude correction	50

Table 2: Articulation score: effect of modulation envelope correction

4 ITERATIVE NOISE REMOVAL

This section describes a new approach to enhancement of reverberant speech which attempts to estimate the reverberant component of the signal using the measured segmental smoothed spectral magnitude characteristics of the room, which are assumed independent of position. This estimate is used in a noise reduction enhancement paradigm rather than the usual deconvolution approach. This assumes that speech segments are independent and uncorrelated — an assumption which obviously is often poor. The degree of correlation will determine the effectiveness of this approach.

Figure 1 compares LPC smoothed spectral magnitudes of corresponding 20ms segments of the three measured positions (offset for clarity). Examination of the smoothed spectrum of these (and other) segments of the three impulse responses at similar temporal displacements shows a high degree of similarity. This similarity appears to correspond with the subjective similarity of reverberation produced in the different positions.

To use the observed similarity in segmental smoothed spectral magnitudes, we segment the room response into frames of size N, so that

$$h(n) = \sum_{m=0}^{M} h_m(n - mN) \quad (5)$$

Then

$$\begin{aligned} x(n) &= s(n) * h(n) \\ &= \textstyle\sum_{m=0}^{M} s(n) * h_m(n - mN) \\ &= \textstyle\sum_{m=0}^{M} s(n - mN) * h_m(n) \end{aligned} \quad (6)$$

The reverberant signal consists of the latter components, i.e. $m > 0$. If the first (or first few) components of this sum can be recovered, we would expect a reduction in perceived reverberation.

We approximate the $h_m(n)$ by their linear predictive coefficient vectors $A_m(n)$ and gain factors g_m:

$$x(n) \approx \sum_{m=0}^{M} g_m\,(s(n - mN) * A_m(n)) \quad (7)$$

From the observed similarity of LPC smoothed spectra, we assume that the A_m and g_m are independent of source or receiver position and approach the problem as an additive noise problem rather than as a convolution. We assume that

$$s(n) * A_M(n) \approx x(n) * A_M(n) \tag{8}$$

We then use an iterative procedure, starting from the 'tail' of the impulse response. The initial signal estimate $\tilde{s}_0$ is set to the reverberant signal. The contribution of each segment of $h(n)$ to the reverberant portion of the signal is estimated and removed by standard noise reduction techniques. Thus (with k a small constant),
for $m = 1$ to $M - k$:

$$\begin{aligned} r_m(n) &= g_{M-m}\left(\tilde{s}_{m-1}(n - mN) * A_{M-m}(n)\right) \\ \tilde{s}_m(n) &= \mathrm{SS}\left(\tilde{s}_{m-1}(n), r_m(n)\right) \end{aligned} \tag{9}$$

where $\mathrm{SS}(p, q)$ denotes the spectral subtraction from signal p of signal q.

Due to phase differences, only a proportion of the estimated signal spectrum is subtracted at each iteration: typically between 0.01 and 0.1 in this case. A value of 0.7 was used for all test set utterances.

Table 3 shows articulation scores from the intelligibility test using speech reverberated in both reference and rotated positions, both enhanced using the parameters of the reference position. The intelligibility test set was derived using a root cepstral subtraction procedure instead of spectral subtraction, as described by Fisher and Sridharan [2]. Figure 2 shows an example of waveforms of the clean and reverberant speech, and the processed speech for two of the positions.

An example of the iterative technique is available for the utterance "chip watch joke" in proceedings CDROM sound files containing reverberant speech [A0595S06.wav] and processed speech [A0595S07.wav]. These show significant reduction in the perceived level of reverberation.

Condition	AS (% correct)
(a) Raw	51
(b) Reference	51
(c) Rotated	50

Table 3: Iterative spectral subtraction Articulation Score *AS* (% correct) for:
(a)raw reverberant speech; iterative spectrally subtracted speech in the (b) reference and (c) rotated positions.

5 CONCLUSIONS

The results obtained show clearly that the waveform deconvolution (inversion) approach is impractical in any situation where either source or receiver is mobile, due to variation of room impulse response with position. Although the modulation envelope deconvolution method has been claimed to be position independent, our results indicate this reduces intelligibility of the speech signal (although the technique might be useful for quality improvement).

The iterative noise reduction method proposed is unaffected by positional change at least over moderate ranges. Although intelligibility was not improved in the tests performed, the lack of degradation is encouraging when compared with previously proposed methods. It should also be noted that the technique was not optimised — for each utterance, a different proportion of reverberant signal estimate should be removed for best results. A higher proportion than was used is beneficial for some utterances, while completely removing the desired speech from others.

A method of determining the optimal proportion for individual utterances is an obvious manner of improving the procedure. For this purpose, the modulation envelope deconvolution method might be of use to provide a constraint. In such a scheme, a small proportion of reverberant signal estimate would be removed repeatedly until the estimated modulation envelope was approximated for each speech segment.

References

[1] Cole D., Moody M., Sridharan S. "Intelligibility of reverberant speech enhanced by inversion of room response", Proceedings of the International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN–94), I, pp.241-244, 1994.

[2] Fisher A., Sridharan S. "Real-time audio enhancement system", Journal of the Audio Engineering Society **43** (7/8), p.634, 1995.

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

[4] Mourjopoulos J.;, Hammond J.K. "Dereverberation of speech signals using an envelope convolution model", IEEE ICASSP-83, pp.1144-7, 1983.

[5] Mourjopoulos J., Clarkson P., Hammond J. "Dereverberation of speech using optimum control", Digital Signal Processing-84, pp.415-419, 1984.

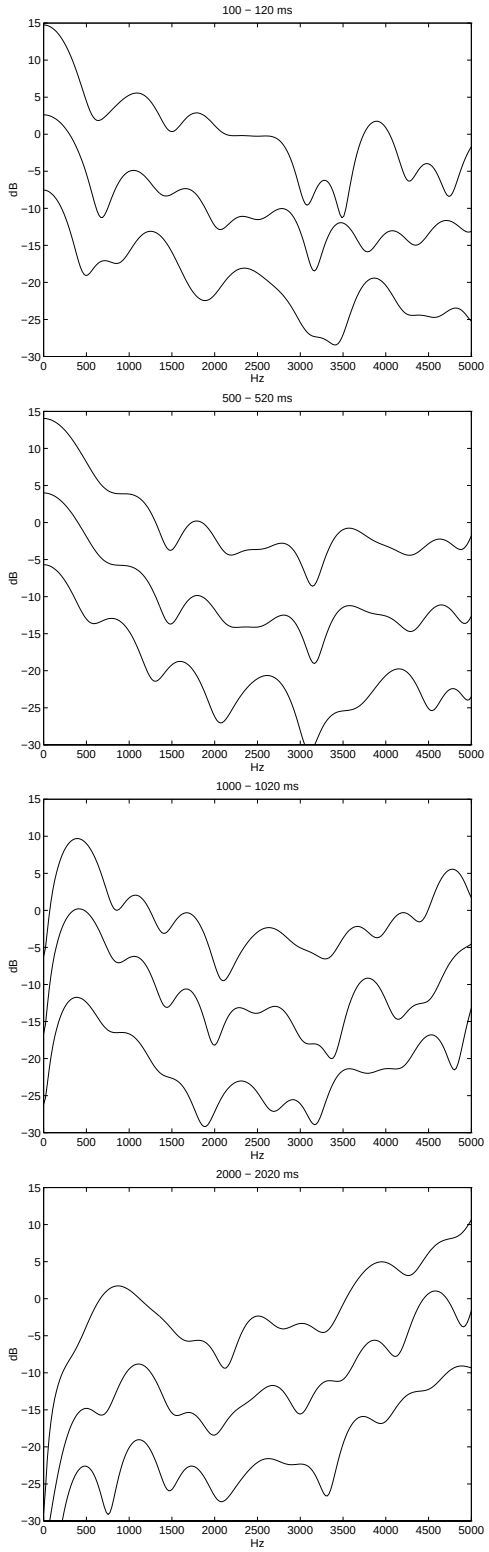

Figure 1: Segmental smoothed spectral magnitudes
upper trace: reference position $h_{ref}(n)$
centre trace: rotated position $h_{rot}(n)$
lower trace: shifted position $h_{sh}(n)$

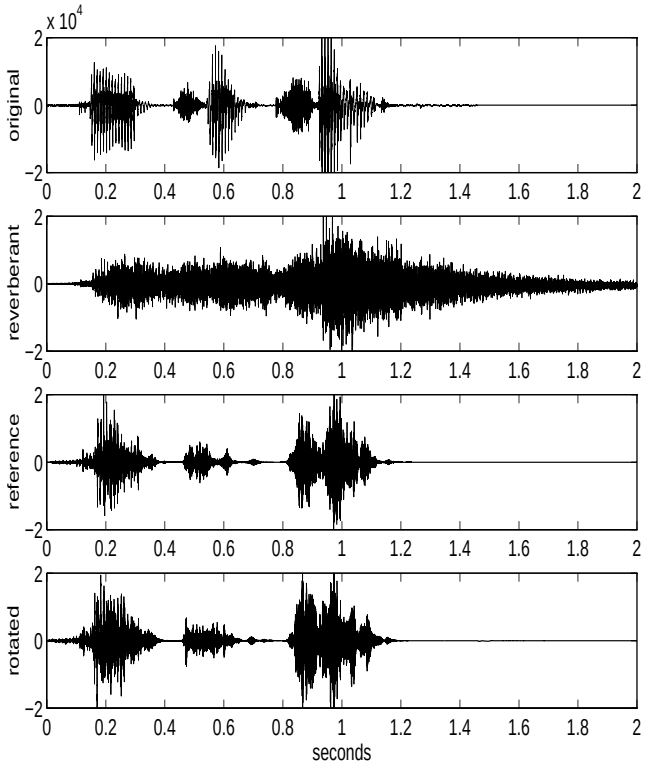

Figure 2: Iterative spectral subtraction waveforms

Original speech; reverberant speech; resultant for reference position; resultant for rotated position.