MEASURING INTELLIGIBILITY OF REVERBERANT SPEECH WITH AND WITHOUT ENHANCEMENT

D. Cole, M. Moody and S. Sridharan

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

ABSTRACT — Three acoustical measures are evaluated as predictors of speech intelligibility for highly reverberant speech both before and after enhancement by inversion of the room impulse response. The predictors are the Speech Transmission Index (STI), the Lochner-Burger signal to noise ratio, and a simple ratio of useful to detrimental energy. Predictions are compared with articulation scores using phonetically balanced (PB) speech material. Intelligibility improvements were consistent with two variants of articulation test, but were not reflected in any predicted value for short inverse filter lengths.

INTRODUCTION

Dereverberation of speech is often investigated for teleconferencing and mobile telephony situations, where multi-microphone techniques may be used. Such methods greatly increase the effectiveness of processing over single input processing. In forensic applications, 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.

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.

Recently, 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 is limited severely by the requirement for a minimum phase characteristic after windowing, and by variation of $h(n)$ with positional changes.

Calculation of the inverse response may be made in two ways, using either cepstral or least-squared error approaches (Mourjopoulos *et al* 1982).

The reported techniques of Mourjopoulos *et al* (1982) and Bees *et al* (1991) have several shortcomings when applied in the forensic situations described.

- Results have been reported only as improvement of direct to reflected signal energy ratios. Intelligibility improvement, which is the requirement for forensics, is not necessarily well correlated.
- Simulated room responses typically of 256 samples at 8 to 10kHz have been used for testing. Forensic applications often involve much more difficult characteristics. High surface reflectivity in spaces of 30 to 40 cubic metres typically give reverberation times of several seconds, with very dense impulse responses. It is easy to demonstrate that an extremely severe exponential weighting function is required to approach a minimum phase characteristic for such responses. Numerical accuracy thus is compromised beyond the initial portion of the impulse response. It follows that only the initial portion of such impulse responses may be estimated using the procedure in Bees *et al* (1991).
- The degree of improvement gained using partial estimates of the response has not been quantified. This is of great interest considering the previous point.
- The effect of variation of source/receiver position is of great importance since it may vary greatly even within the reverberation time of the room (Mourjopoulos 1985).

We have previously reported the effect of using truncated estimates on intelligibility (Cole *et al* 1994). This paper extends that study by comparing the intelligibility predictions of several measures with the articulation scores obtained.

METHOD — ROOM MODELLING

The impulse response of a bare room (of about $40m^3$, with reverberation time of about 3 seconds) was measured using a chirp signal, as outlined in Cole *et al* (1994). This response was used to artificially reverberate clean speech utterances for articulation testing, and to design various inverse room filters. These filters were designed from the initial portion (of varying length) of the measured response, simulating perfect estimation of that portion.

A straightforward application of recursive least squares techniques was used to design the inverse filters. The room impulse response estimate was fed to a two input recursive least squares filter, with an impulse supplied as the second (reference) signal. As the room characteristic generally is mixed phase, this reference impulse must be delayed to allow inversion of maximum phase components. These two signals may be passed through the filter as many times as necessary for the desired degree of convergence. Given a good impulse response estimate, each iteration gives improved quality (and intelligibility) of the dereverberated speech.

The five inverse filters were designed from truncated impulse responses of lengths 1024, 2048, 4096, 8192 and 32768 (93, 186, 372, 743 and 2972 msec respectively). Iterations continued until reduction of total squared error reduced by less than 0.001 between iterations. The inverse filters used were twice the length of the truncated response for each case. The first half of each inverse filter provided the acausal maximum phase inversion, thus introducing a delay equal to the length of the truncated impulse response.

Six conditions were tested for intelligibility: the unprocessed reverberated speech, and the results of deconvolving the reverberated speech with the five inverse filters of differing lengths.

METHOD - INTELLIGIBILITY TESTING

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. Each subject was presented with fifty words for each condition, with each word enclosed in the carrier phrase "Word [number] is [word]." Thus a total of 600 trials were conducted for each filter.

After analysis of the results of the preceding tests, another test was designed to overcome apparent deficiencies in the testing method. The revised procedure used groups of three PB words uttered sequentially (but quite distinctly, to avoid co-articulation effects.) A carrier phrase was not used. Each listener recorded all three words, with fifty such tests for each condition, giving 1200 test words for each filter. The effects of the longest two inverse filters were not evaluated in this test.

METHOD - PREDICTORS

The three predictors calculated were the Speech Transmission Index (STI), the Lochner-Burger signal to noise ratio (C_{95}), and the ratio of useful to detrimental energy (U_{80}) as evaluated by Bradley (1986) for various rooms.

The STI was calculated using the modulation transfer function in octave bands and with weightings as suggested by Houtgast and Steeneken (1980). Schroeder's theoretical formula for STI calculation (Schroeder 1981) was also used.

The Lochner-Burger C_{95} measure is a form of early/late ratio, with a time- and amplitude-dependent weighting to determine the contribution of early arriving reflections (ie, within the first 95ms) to the useful energy. An identical formula and weighting to that of Bradley (1986) was used:

$$C_{95} = 10 \log \frac{\int_0^{0.095} \alpha(t)\, h^2(t)\, dt}{\int_{0.095}^{\infty} h^2(t)\, dt} \tag{2}$$

where:

$$\alpha(t) = 2.3 - 0.6 \left(\frac{h(t)}{h(0)} \right)^{0.7} - t \left(0.0248 - 0.00177 \left(\frac{h(t)}{h(0)} \right)^{1.35} \right) \tag{3}$$

and $\alpha(t)$ is further limited such that $0 \leq \alpha(t) \leq 1$

The U_{80} useful-detrimental ratio was calculated as the ratio of energy arriving in the first 80ms to the energy arriving after. In its complete form, uncorrelated additive noise is also taken into account. In this case, this component is zero and the expression for U_{80} is similar to that for C_{95}, with $\alpha(t) = 1$ and an 80ms integration limit. Bradley (1986) found this to be as good a predictor as the more complex ratios in his study of auditoria.

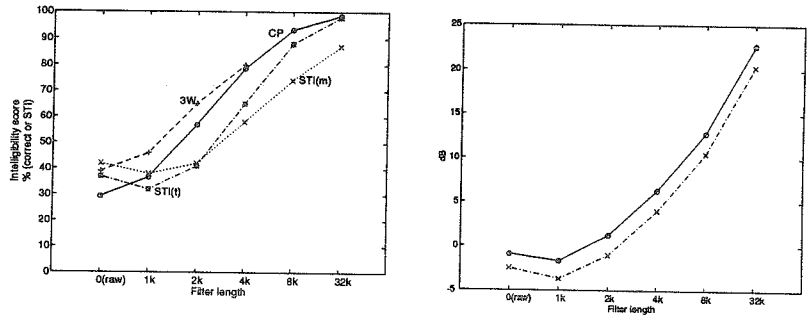
A phenomenon encountered in this study which does not occur in real, physical situations is the arrival of many 'pre-echoes' well before the 'direct' sound. This occurs due to the delay introduced to enable the 'acausal' inversion of maximum phase components. For the purpose of predictor calculation, these pre-echoes were ignored (except for the STI measurement). It was noted that other options tried, which accounted for these in some way, gave similar trends as the results presented below.

RESULTS

The effects of room response inversion for different filter lengths has already been reported in Cole (1994). The interesting and unexpected result is the improvement in intelligibility with even the shortest (93ms) filter. This has been attributed to whitening of the room response.

Several interesting points were noted during articulation testing. The initial test performed was the carrier phrase test, and analysis indicated problems due to the phonetically unbalanced carrier phrase. In particular, the preceding /z/ of 'is' often was incorporated into the response as either a leading or trailing phoneme. (This is noted in Cole (1994).) Inter-subject variance was increased by a tendency by some respondents to assume co-articulation effects, or to adjust their perceived responses to fit a 'real' word. A better choice of test material to avoid this might be to include equal numbers of nonsense CVC syllables in the PB word lists.

The articulation scores and predictor values obtained are presented in Figure 1. It is obvious from this that the predictors have similar characteristics. This agrees with the findings of Bradley (1986). They correlate well with articulation scores when reverberation is the sole degradation mechanism, but all fail to predict well the effect of the shorter filters. We conclude that the predictors may be considered accurate where there is only one source of degradation. None of the predictors successfully balance the effects of spectral distortion and temporal masking phenomena present in our situation. The similarity between the C_{95} and U_{80} predictors also agrees with the results of Bradley. The α weighting factor used in the calculation of C_{95} in this study is identical to that reportedly used by Bradley. Analysis of this factor shows that it simulates the results of Lochner and Burger over a limited amplitude range. This might explain the similarity to the unweighted U_{80} measure.



(a) (b)

Figure 1: Intelligibility test results
(a) CP: *carrier phrase test*; 3W: *three word test*; STI(m): *measured STI*; STI(t): *theoretical STI*
(b) upper: C_{95}; lower: U_{80}

ACKNOWLEDGEMENTS

This work was assisted by grants from the National Institute of Forensic Science, the Queensland Police Service and a QUT Research and Development Encouragement Award.

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.

Houtgast H.J.M., Steeneken T. (1980) "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Am. **67** (1), 318–326.

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

Lochner J.P.A., Burger J.F. (1964) "The influence of reflections on auditorium acoustics", J. Snd. Vib. **1** (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.

VECTOR QUANTIZATION OF DCT COMPONENTS FOR SPEECH CODING

Jianwei Miao
School of Electrical and Computer Engineering
Georgia Institute of Technology

ABSTRACT - In this study a system of design algorithms and experimental results was investigated for vector quantization of discrete cosine transform components in speech coding. The encoding and decoding systems of vector quantization of discrete cosine transform components were developed for digital speech coding. A training sequence consisting of 200,000 speech samples sampled at 8 kHz with 16 bits signed integers was used to design the codebooks. The codebooks were designed using the K-means algorithm. The implementation of a discrete cosine transform vector quantizer used a two-codebook structure, providing different codes for different real component vectors corresponding to different frequency bands. This is a form of subband coding and yields a means of optimizing bit allocations among the subcodes as well as produces a good quality speech at low bit rates. The experimental results of implementing the encoding and decoding systems showed a good quality speech at 7111 BPS with 15.61 dB in signal-to-quantization noise ratio.

INTRODUCTION

The goal of selecting the transformation from the time domain to the frequency domain for digital speech coding is to obtain uncorrelated spectral samples. The Karhunen-Loéve transform (KLT), in this sense, is optimum which it yields uncorrelation of spectral values. However, this KLT is difficult to calculate in general. For instance, given a frame of N points, $O(N^4)$ flops are necessary to compute the KLT (Wintz, 1972). Furthermore, the whole autocorrelation matrix of the frame must be sent to the receiver as side information in order that the transform may be inverted. The discrete Fourier transform (DFT) and discrete cosine transform (DCT) are viable alternatives. The DFT and DCT are not optimum which they do not decorrelate the data, but they have the advantage of not depending on the data statistics and will approximately decorrelate the components when the block length is sufficiently long (Pearl, 1973, Ahmed, et al., 1974, Chang, et al., 1987). Of these two, the DCT yields good performance compared with the KLT and is generally used in practice (Deller, et al., 1993).

The application of vector quantization (VQ) techniques in the DCT components has recently emerged as a powerful means for digital speech compression. Because the DCT of the digital speech data prior to quantization has three potential advantages. First, each sample in the DCT domain depends on many samples in the original domain. Second, DCT vector quantization of the digital speech waveforms supplies apparently better subjective quality than ordinary vector quantization of the digital speech waveforms using codes of comparable complexity. Third, quantization noise can be shaped by allocating bits to the DCT components according to a 'perceptual criterion'.

The purpose of this paper is to investigate the VQ of DCT components for digital speech coding. The implementation of a DCT vector quantizer used a two-codebook structure which provided