A NETWORK SPEECH ECHO CANCELLER WITH COMFORT NOISE

D.J.Jones, S.D.Watson* ,K.G.Evans*, B.M.G.Cheetham* and R.A.Reeves#.*
*Department of Electrical Engineering, The University of Liverpool, Liverpool, L69 3BX, UK.
#BT Laboratories, Martlesham Heath, Ipswich, IP5 3RE.

Tel: +44 (0)151 708-7724 E-mail: davej@liv.ac.uk

ABSTRACT

This paper describes a proposed comfort noise system for a network echo canceller. In this system, any residual echo is suppressed using a single threshold centre-clipper, but instead of transmitting silence to the far-end of the network, a synthetic version of the background sounds is sent. This masks any ‘noise modulation’ or ‘noise pumping’ that may otherwise occur. The background sounds are characterised using linear prediction. Periods when only background sounds are present are identified by a modified GSM Voice Activity Detector (VAD). Informal listening tests have shown that this ‘synthetic background’ is preferable to the transmission of silence or pseudo-random noise that is not spectrally shaped to match the original background.

1. INTRODUCTION

For telephone calls with a long round-trip delay, the operation of adaptive echo cancellers has a large effect on the perceived quality of the call. In order to cancel any echo, most network echo cancellers use an adaptive filter whose coefficients are adjusted to generate a replica echo which is subtracted from the actual echo, as in figure 1. However, in practice the echo cannot be entirely cancelled by the adaptive filter due to factors such as non-linearities in the echo path, A/μ-Law companding and, background sounds from the near-end of the network. Generally, in the absence of near-end background sounds, the maximum echo attenuation, or Echo Return Loss Enhancement (ERLE), is determined by the companding process. Under these conditions, a large component of the residual echo is quantisation noise which, if not removed, would be disturbing to the far-end talker.

The quality of performance of the canceller must be maintained even in the presence of a high level of background noise, for example when a telephone is used in a noisy office or when a mobile telephone is used in a car. It is likely that the ERLE will be limited by the background sounds which may have a larger power than the quantisation noise. The residual echo will still be audible and just as disturbing as quantisation noise in the absence of near-end talker.

In order to reduce the residual echo power to acceptable levels [1] a non-linear processor (NLP), or centre-clipper, may be used. A centre-clipper attenuates signals whose amplitude is less than a threshold, which should ideally be set to the peak amplitude of the residual echo. In this way, the residual is suppressed but large amplitude signals are allowed to pass. The clipper is disabled in the absence of echo and when near-end speech is present, and this allows signals to pass without distortion. When the clipper is operating, the far-end talker will hear silence or the idle channel noise of the digital long distance circuit. When switched out, the far-end talker will hear the near-end background sounds plus the idle channel noise of the analogue near-end. This switching between different noise levels is known as ‘noise modulation’ or ‘noise pumping’ and is undesirable. Clearly, the characteristics of the centre-clipper have the potential to severely degrade the perceived call quality, particularly in the presence of near-end background sounds.

2. COMFORT NOISE

Some echo cancellers attempt to mask noise modulation by implementing comfort noise that is similar to noise-matching in DCME with Digital Speech Interpolation (DSI) [2,3]. In this scheme, pseudo-random noise matched to the power of the background is injected during periods of silence when the centre-clipper is operating. However, informal testing has shown that although much less disturbing than noise modulation, the switching between comfort noise and background noises is still disturbing and does not sound realistic.

A comfort noise algorithm which addresses these limitations has been developed. In this system, shown in figure 1, a synthetic version of the background sounds is generated and added to the clipper output when required. The background sounds are characterised using the discrete all-pole speech model which enables the spectrum of the comfort noise to be matched to the actual background sounds. A voice activity detector (VAD) is used to indicate when only background sounds from the near-end are present and during these periods the background characteristics may be modelled and stored as a sequence of parameter sets for use in generation of comfort noise.

The following sections discuss the operation of the spectral matching process, the VAD and comfort noise generation mechanism.

3. SPECTRAL MODELLING

Although there are many different techniques for spectral estimation, the comfort noise scheme described here assumes that the background sounds can be modelled using a discrete all-pole model. Linear Prediction [4] is used to obtain sets of coefficients that represent the spectral envelope of the background and Durbin's algorithm [4] is used to compute LPC parameters from autocorrelation coefficients derived in frames of background sounds.

It is to be expected that the all-pole model will be a good approximation to the spectral envelope if the background sounds are speech-like. During voiced periods the poles will relate to the formant frequencies and this results in a spectral envelope that is a close match at these frequencies. During unvoiced periods the formant structure is no longer present but the poles are positioned so that the spectral envelope is still a good match to the frequency response of the vocal tract.

Other types of background sound such as car or babble noise, are likely to have a spectral structure that is more similar to unvoiced speech than voiced speech. Car noise, for example, does not exhibit the 'peaky' spectral nature of voiced speech. It is thus reasonable to assume that the all pole model will still give a good representation of the background sounds.

In addition to the spectral shape, the comfort noise should ideally exhibit the same spectral variation with time for it to be realistic. This can be achieved by storing a sequence of coefficient sets that represent the time variation of the background spectrum.

At intervals of 10ms, the linear prediction coefficients of the background are calculated by the VAD, from the average autocorrelations of the last four frames. When the VAD indicates background noise only, these coefficients are added to a cyclic buffer. As new sets are added, the oldest in the buffer are replaced so the spectrum of the comfort noise can approximate the characteristics of the latest background sounds, in terms of both spectral shape and time variation.

4. VOICE ACTIVITY DETECTOR (VAD)

A Voice Activity Detector (VAD) is used to indicate the presence of speech and echo from the near-end. This is important as the comfort noise must 'mimic' the background sounds rather than any speech that may be present.

The VAD that is used is based on the GSM voice activity detector, which has been shown to give accurate detection of speech in noisy environments [5]. It consists of two separate detection units, the primary and secondary VAD's. Using an inverse filter, the primary VAD attempts to filter out any background noise which is assumed to be stationary and non-periodic, and the energy of the filtered signal is compared with an adaptive threshold to decide whether speech is present. This provides robust detection in the presence of high background noise levels. The secondary VAD decides when to update the coefficients of the inverse filter and adapt the primary VAD threshold and this occurs when the input is classified as stationary and non-periodic. The secondary VAD cannot be used in isolation because the stationarity and periodicity tests are very strict and hence a relatively small portion of the background noise will be detected.

Normally on initialisation, the VAD takes some time to adapt its threshold and thus give accurate discrimination between speech and non-speech. For some background sounds this takes only a few seconds, but for sounds such a speech, the adaption can take much longer. In order to ensure accurate operation at start-up, the VAD is modified to assume that, for the first 0.5 seconds after call initialisation, only near-end background sounds will be present. During this period the VAD is forced to adapt the primary threshold and indicate that background noise is present.

5. COMFORT NOISE GENERATION

The comfort noise is generated by a synthesis filter using a random Gaussian excitation and a set of prediction coefficients. The coefficient sets are retrieved from the cyclic buffer in reverse order and when all have been played back, they are repeated. After synthesis, the comfort noise power is scaled to match that of the background. Normally, a predictive speech coder would use an excitation that is an impulse train/random noise combination to generate voiced/unvoiced sounds. The use of such an excitation in this application is undesirable because the comfort noise would be identical to past background sounds. The far-end talker would hear repeated background when the clipper is active. For some types of background such as speech, the repetition will be more disturbing than for other kinds of background, such as for car noise. In all cases, the switching between real and synthetic sounds is likely to be very obvious. It is not desirable for the comfort noise to sound identical to the original background, just similar, and this may be achieved using a random excitation.

6. TESTING

The performance of the comfort noise algorithm has been evaluated using a high level language computer simulation of the system shown in figure 1. However, the adaptive filter was simulated rather than using an adaption algorithm such as the LMS algorithm [6]. This is more convenient because there is no delay in waiting for the filter to converge to its steady state level and the misadjustment can be set independently of step-size and near-end background noise power. A simple energy based double-talk detector was used to enable the centre-clipper when echo is present and disable it otherwise. Testing was carried out using several different noise types: i) car noise, ii) babble noise and multi-speaker noise, at different power levels, calibrated relative to the near-end speaker.

Figures 2 and 3 are spectrograms of the residual echo r(n), and the processed residual $r_{NLP}(n)$ respectively. The light areas are of high energy due to residual echo (RE) or near-end talker (NE) and darker areas are of lower energy due to car noise at 0dB relative to the near-end talker. The clipper is active during periods of echo and hence the echo is removed and replaced by the synthetic background sound. It can be seen that the comfort noise has the same spectral shape as the actual background. There are no obvious discontinuities between the real and synthetic backgrounds. For other types of background sound, the spectrograms (not shown here) exhibit similar behaviour.

Informal listening tests suggest that the addition of the synthetic background with its time variation is more natural than when constant unshaped pseudo-random noise is injected. For ‘unvoiced’ environmental noises such as car or babble noise, the synthetic background is almost identical to the original background. The far-end talker will not be aware of switching between comfort and real background unless specifically listening for it. When the background consists of ‘voiced’ type sounds such as multi-speaker noise, the comfort noise is still speech-like even though it contains no periodic components. It also has similar spectral variation to the actual background even though the time varying linear prediction coefficients recreate the spectral variation of past background sounds. In both cases the switching between comfort noise and real background is much less disturbing than noise modulation and hence the suppression of the residual is more acceptable.

If listened to in isolation, ‘looping’ of the comfort noise is sometimes audible but is dependent on the type of background and number of coefficient sets captured. For example looping in car noise is much less conspicuous than in multi-speaker noise because the spectral envelope of car noise changes much less than that of multi-speaker noise.

7. CONCLUSIONS

In conclusion, the advantages of this comfort noise algorithm are, firstly, that there will be complete suppression of the residual echo, assuming the

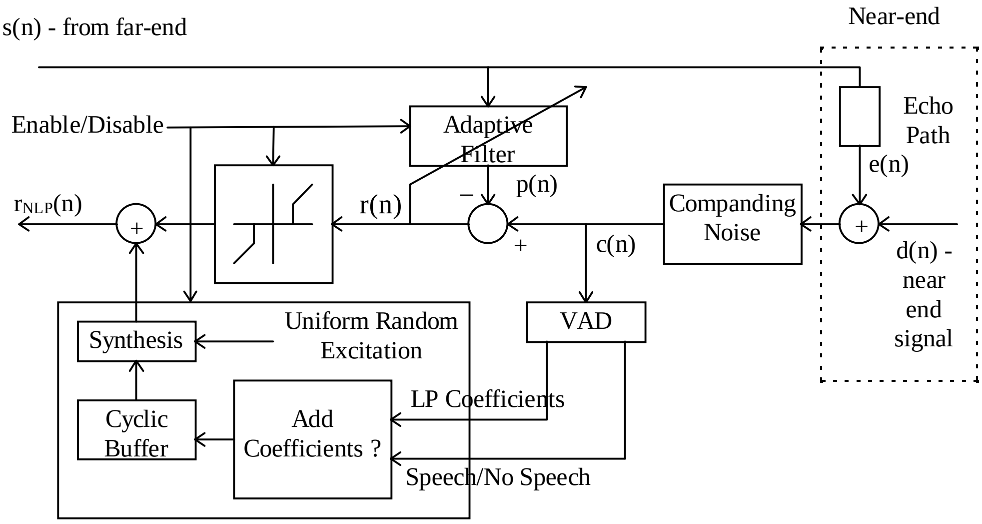

Figure 1

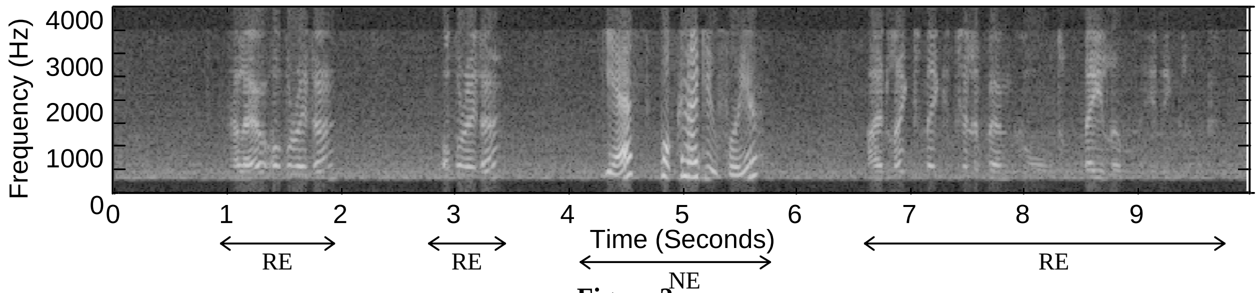

Figure 2

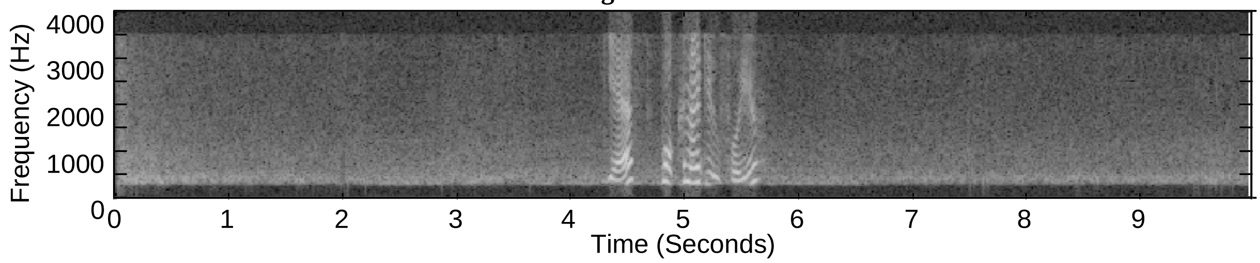

Figure 3 (RE, Residual Echo; NE, Near-End Signal)

clipper threshold is set correctly, and secondly, the background sounds heard at the far-end will be more realistic during periods of residual echo when the clipper is operating, resulting in negligible noise modulation. The system is currently being incorporated into a DSP based real-time echo canceller so that it may be tested for conformance with the standards and to permit real-time subjective testing.

8. REFERENCES

[1] ITU-T Recommendation G.165, "Echo Cancellers"

[2] T. L. Barto, "New Considerations for Echo Control in the Evolving Worldwide Telecommunications Network", Tellabs Inc., Proceedings Telecom 1991

[3] ITU-T Recommendation G.763, "Digital Circuit Multiplication Equipment using 32kbit/s ADPCM and DSI"

[4] J. Makhoul, "Linear Prediction: A Tutorial Review", Proceedings of the IEEE, April 1975.

[5] S. D. Watson, B. M. G. Cheetham, W. T. K. Wong, P. A. Barrett and A. V. Lewis, "A Voice Activity Detector for the ITU-T 8kbit/s Speech Coding Standard G.729", Eurospeech 1997.

[6] B. Widrow, J. M. McCool, M. G. Larimore and C. R. Johnson, "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter", Proceedings of the IEEE, vol. 64, no. 8, August 1976.