A SIGNAL SPECTRUM SUBTRACTION ALGORITHM FOR ROBUST SPEECH RECOGNITION

Michael J. Carey[1], Graham D. Tattersall[2] and Eluned S. Parris[1]

[1]Ensigma Ltd., Turing House, Station Road, Chepstow, Monmouthshire, U.K
[2]Snape Signals Research, New House Friston Suffolk U.K.
{michael eluned}@ensigma.com, graham@snapesignals.co.uk

ABSTRACT

The subject of this paper is the problem of mitigating the effect of an interferer such as in-car-entertainment ICE playing during the speech recognition process. In this case we have a known non-stationary interferer which is added to the speech having passed through an unknown acoustic channel. This problem has been addressed to date using echo cancellers based on adaptive filters, however we proposed a frequency domain technique solution based on channel identification followed by spectral subtraction. We provide the results of experiments which show that about 10dB of cancellation is possible with loud background music. While this is insufficient cancellation to satisfy a human listener it will prove sufficient to substantially improve the performance of a speech recogniser when this type of interference is present.

1. INTRODUCTION

If speech recognition is to find widespread application the problem of maintaining performance in the presence of interfering signals must be solved. To date attention has bee directed towards mitigating the effect of quasi-stationary noise such as telephone channel noise or car noise on recogniser accuracy and several techniques have been proposed to deal with these. These techniques include spectral subtraction[1] Weiner filtering[2] and parallel model combination[3]. Each of these techniques works in the spectral domain.

There are however other sources of interference in the acoustic environment. One of these is sound generated by an electronic device under the control of the user. This could be a radio, CD tape or telephone message, we generically refer to these as In Car Entertainment (ICE) signals. These interferers could be present when the user wishes to say a voice command. For example the radio may be playing in a car when the user wants to use voice control of the navigation system or the radio itself. In this case the original interfering signal is assumed to be known and accessible but has passed through an unknown acoustic path between the loudspeaker and the microphone which is characterised by some unknown impulse response. The main approach used until now has been the use of acoustic echo cancellers [4] based on time domain adaptive filters. While these may be appropriate, adaptive filtering suffers from a number of disadvantages. These include the high computational requirement and the slow convergence of the algorithms.

In this paper we propose an alternative approach based on spectral domain processing of the interferer. In order to do this we assume that:

1) The phase of the interferer is not required at the recogniser,
2) The degree of interference rejection required is smaller than that needed to satisfy a human listener in a speech enhancement application,
3) The interfering signal may be mono or stereo,
4) Each frame of the interferer's energy is concentrated into a single analysis frame.

We justify these assumptions by noting that recognition feature sets such as the cepstra do not contain phase information. A human listener is sensitive to levels of interference 40db below the level of the wanted signal while a speech recogniser can operate well with a 15db signal to noise ratio. Since observations we have made indicate that the worst case signal to noise ratio in the presence of a high level interferer is about 5dB the reduction in the level of the interferer of about 10dB is required.

The remainder of this paper is organised as follows. In section two we outline the algorithm while in section three we describe the system we propose for interference cancellation. In section four we show results of our initial experiments.

2. THEORY

2.1 Problem Definition

Figure 1 shows a simple situation in which left and right stereo ICE signals, $L(j\omega)$ and $R(j\omega)$ are transmitted from separate loudspeakers. These signals are added to the wanted speech signal, $S(j\omega)$ at the in-car microphone. Perfect cancellation of the unwanted ICE signals could in principle be achieved given knowledge of the left and right acoustic transfer functions, $H_{AR}(j\omega)$, $H_{AL}(j\omega)$, and the source stereo signals $L(j\omega)$ and $R(j\omega)$. Although the source signals $L(j\omega)$ and $R(j\omega)$ are readily available, the acoustic transfer functions must be estimated. This is the core problem that must be solved to make ICE cancellation possible.

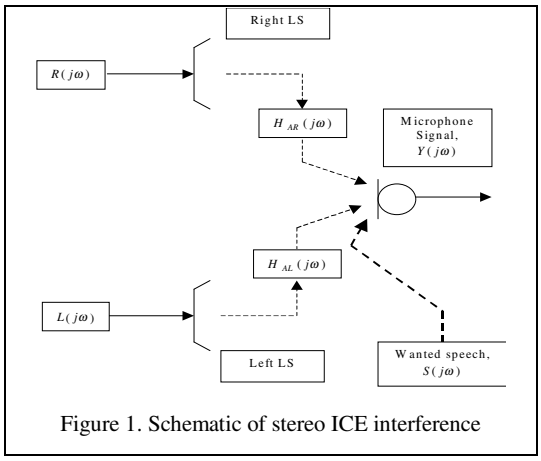

Figure 1. Schematic of stereo ICE interference

2.2 Channel Estimation

A simple approach to the estimation of the acoustic transfer functions is to find the long term ratio of microphone signal spectrum to each of the source stereo signals. The following equations show this process for the right acoustic channel. A similar set of equations can be written for the left channel.

$$\hat{H}_{AR}(j\omega) = \frac{Y(j\omega)}{R(j\omega)} \quad ..1$$

The spectrum of the microphone signal is:

$$Y(j\omega) = H_{AR}(j\omega).R(j\omega) + H_{AL}(j\omega).L(j\omega) + S(j\omega) \quad ..2$$

So, the transfer function estimate for the right acoustic channel becomes:

$$\hat{H}_{AR}(j\omega) = H_{AR}(j\omega) + H_{AL}(j\omega).\frac{L(j\omega)}{R(j\omega)} + \frac{S(j\omega)}{R(j\omega)} \quad ..3$$

Three conclusions are drawn from equation 3:

- If $L(j\omega)$, $R(j\omega)$, and $S(j\omega)$ are all uncorrelated, a correct estimate of the channel response will be obtained because the second and third terms in the expression will have long term averages of zero.
- If $L(j\omega)$ and $R(j\omega)$ are completely correlated as in a mono transmission, whilst still being completely uncorrelated with $S(j\omega)$, the individual left and right channel transfer functions cannot be uniquely determined, but a composite estimate which contains terms due to both left and right channels can be obtained. This is sufficient for perfect cancellation of the mono ICE signal at the microphone.
- If $L(j\omega)$ and $R(j\omega)$ are partially correlated, the left and right acoustic channels cannot be unambiguously estimated. However, if $L(j\omega)$ and $R(j\omega)$ occupy different spectral regions or if $l(t)$ and $r(t)$ have periods when one has low energy whilst the other has high energy, it may still be possible to make useful estimates of left and right channels for the purposes of cancellation.

2.3 The ICE Cancellation Algorithm

Frequency domain estimation of the left and right acoustic channel responses using equation (3) is the basis of the ICE cancellation method presented in this paper. As noted earlier, cancellation for the purpose of speech recognition only requires an estimate of the *magnitude* of the speech spectrum because the MFCC feature vector used by the recognition system is based on magnitude spectra. An estimate of the wanted speech magnitude spectrum can be obtained by subtracting estimates of the magnitude spectra of the ICE signals at the microphone.

$$\hat{S}^2(\omega) = Y^2(\omega) - \hat{H}^2_{AR}.R^2(\omega) - \hat{H}^2_{AL}.L^2(\omega) \quad ..4$$

The estimates of the acoustic channel power transfer functions can be made using equation (5). There are two problems with this approach. The first, already discussed, is the problem caused by left-right signal correlation. This can be addressed by using an iterative approach, coupled with time and frequency dimension smoothing of the estimates of the channel responses The second problem arises because the phase information in the channel response is ignored. In reality the phase characteristic encodes a frequency dependent delay spread associated with the acoustic transfer functions. In a car the minimum is about 3ms. The delay spread should be compensated when making the channel estimate using (5). However, this is unnecessary if the spectral evaluation is done using a FFT with block length much greater than the channel delay.

$$\hat{H}^2_{AR}(\omega) = H^2_{AR}(\omega) + H^2_{AL}(\omega).\frac{L^2(\omega)}{R^2(\omega)} + \frac{S^2(\omega)}{R^2(\omega)} \quad ..5$$

A practical form of the ICE cancellation therefore has the following steps:

1. Initialise estimates of the magnitudes of left and right channel transfer functions. $\overline{H}^2_{AR}(\omega) = \overline{H}^2_{AL}(\omega) = 0$.
2. Initialise estimates of the magnitudes of left and right channel interference at the microphone. $C^2_{R,n-1}(\omega) = C^2_{L,n-1}(\omega) = 0$.
3. Make new estimates of the left and right channel interference at the microphone:
   $C^2_{L,n}(\omega) = Y^2_n(\omega) - C^2_{R,n-1}(\omega)$
   $C^2_{R,n}(\omega) = Y^2_n(\omega) - C^2_{L,n-1}(\omega)$
4. Make rough estimates of the left and right channel transfer functions:
   $$\hat{H}^2_{AR,n}(\omega) = \frac{C^2_{R,n}(\omega)}{R^2_n(\omega)}$$
   $$\hat{H}^2_{AL,n}(\omega) = \frac{C^2_{L,n}(\omega)}{L^2_n(\omega)}$$
5. Smooth the rough estimates of the channel transfer function both in the time and frequency dimensions. Time

smoothing is done with a first order recursive filter with time constant of several hundred milliseconds. Frequency smoothing is done with an FIR filter, $f(\omega)$, with triangular impulse response covering about 300Hz. Thus, time smoothing for the right channel is: $\overline{H}^2_{AR,n} = \beta.\overline{H}^2_{AR,n-1} + (1-\beta).\hat{H}^2_{AR,n}$.

Similarly, frequency smoothing for the right channel is:

$$\tilde{H}^2_{AR,n} = f(\omega) * \overline{H}^2_{AR,n}(\omega)$$

The algorithm can be further refined in three ways to deal with the problem highlighted by equation (3) concerning the correlation of the left and right channel signals

- Updating of the recusive filter providing the smoothed channel estimate can be inhibited unless the energy of one channel greatly exceeds the energy of the other channel.
- Updating of the recursively smoothed channel estimate at particular frequencies can be inhibited unless the energy at that frequency in one channel greatly exceeds the energy at that frequency in the other channel.
- Evaluate the coherence function between the left and right channel signals. Then use the inverse magnitude of the coherence at each frequency as a weighting on the amount by which the estimates of the channel responses are updated at that frequency.

3. SYSTEM DESCRIPTION

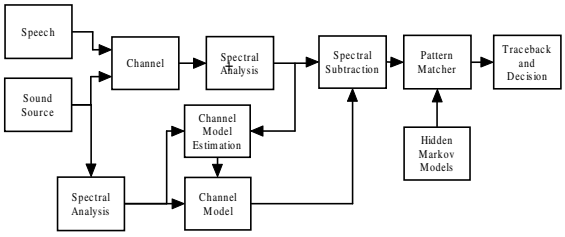

Figure 2 Interference Cancellation Using Spectral Subtraction

The proposed system is shown in Figure 2. When the recognition system is not in use the sound from the ICE source is simultaneously spectrally analysed before and after transmission through the channel. This is carried out at a 16ms frame rate using a 256 point FFT. The two signals are then used to produce an estimate of the transfer functions for the right and left paths. When the recognition system is required the estimate is frozen for the duration of the recognition process. This estimate is then used to perform frequency domain filtering on the source to approximate to the ICE contribution to the signal at the microphone. Spectral subtraction is then used to recover the speech signal from the microphone signal. The recovered signal is then passed to the pattern matcher in the speech recogniser.

Since the algorithm is frame rather than sample based the computational complexity is low. The main computation is required for the FFT which requires NlogN computations per frame for each channel. This is only about 250k computations per second while the simplest form of adaptive filter requires 3N computations per sample. For a echo tail length of 32ms, 256 samples, this equates to more than 18M operations per second. While more efficient adaptive filter systems such as those using sub-band filtering have been described the computation required is still much higher than the spectral subtraction approach.

4. EXPERIMENTS

To allow for comparison between the original signal and the cancelled signal the test data was constructed by recording the speech and interferer separately in the same car environment and then adding the two signals. In both cases the interfering music is a stereo signal. The results of our first experiments are shown in Figures 3 to 6. In Figure 3 we see the microphone signal prior to cancellation when pop music has been added to speech. In this case the peak segmental speech and interferer levels are the same. This is a highly pessimistic way of estimating signal to noise ratio as the amplitude variability of the speech signal is higher than that of the music which exceeds the speech for a considerable part of the example.

The third trace shows the effect of the ICE interferer on the inter cepstral distance between the original speech and the speech plus interferer. These were the frame by frame sum of the squared distance between each of the cepstra, normalised to the frame by frame squared value of the wanted speech cepstra. The second trace shows the recovered speech. This was produced by an inverse transformation on signal after spectral subtraction. The interfering signal has clearly been reduced. This is confirmed by the fourth trace which again shows the normalised squared cepstral distances, but after spectral subtraction. Comparing traces three and four we see that the recovered speech cepstra are less distorted than that with the interferer.

Figure four show similar results for pop music at a better signal to noise ratio of 10dB. While figures five and six demonstrate corresponding results for opera. In each case a useful measure of improvement has been achieved.

5. REFERENCES

[1] P. Lockwood and J. Boudy, "Non-linear Spectral Subtraction and Hidden Markov Models for Robust Speech Recognition in Cars", Proc. ICASSP 92 265-268.

[2] B.P. Milner, "Speech Recognition in Adverse Environments" Ph D. Thesis University of East Anglia 1992

[3] M.J.F Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", Computer, Speech and Language, Vol. 9, No. 4,1995

[4] M. Shozakai, S Nakamura, and K. Shikado, "Robust Speech Recognition in Car Environments", Proc. ICASSP 1998, pp 269-272

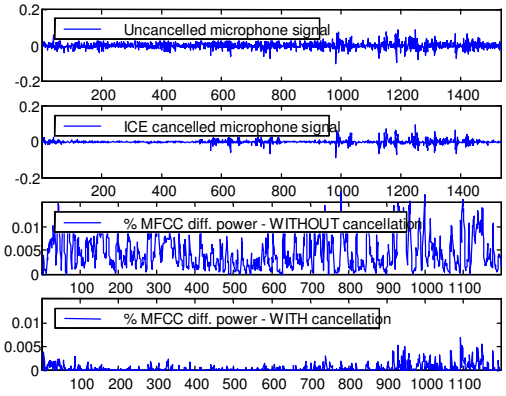

Fig 3: Speech with interfering **pop** music at **0dB** signal to interference ratio. a) uncancelled microphone signal; b) cancelled microphone signal; c) normalised mean square MFCC perturbation **WITHOUT** cancellation. d) normalised mean square MFCC perturbation **WITH** cancellation.

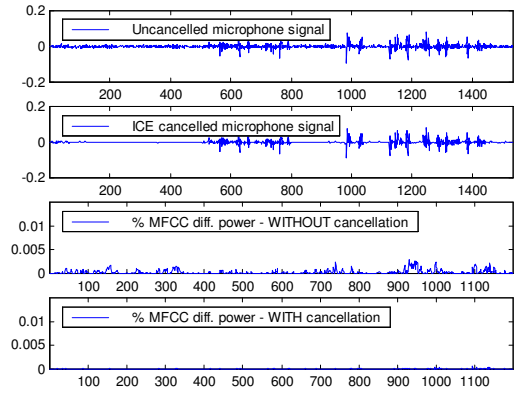

Fig 4: Speech with interfering **pop** music at **10dB** signal to interference ratio. a) uncancelled microphone signal; b) cancelled microphone signal; c) normalised mean square MFCC perturbation **WITHOUT** cancellation. d) normalised mean square MFCC perturbation **WITH** cancellation

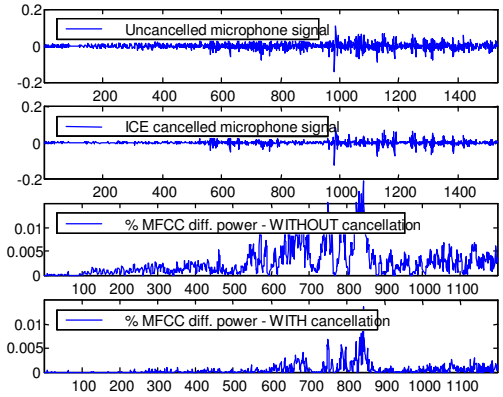

Fig 5: Speech with interfering **opera** music at **0dB** signal to interference ratio. a) uncancelled microphone signal; b) cancelled microphone signal; c) normalised mean square MFCC perturbation **WITHOUT** cancellation. d) normalised mean square MFCC perturbation **WITH** cancellation.

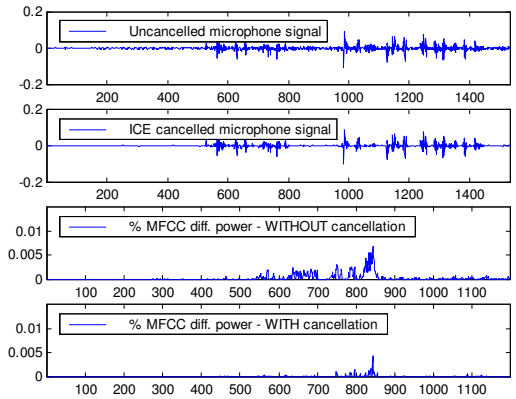

Fig 6: Speech with interfering **opera** music at **10dB** signal to interference ratio. a) uncancelled microphone signal; b) cancelled microphone signal; c) normalised mean square MFCC perturbation **WITHOUT** cancellation. d) normalised mean square MFCC perturbation **WITH** cancellation.