ADAPTIVE NOISE REDUCTION TECHNIQUES FOR SPEECH RECOGNITION IN TELECOMMUNICATIONS ENVIRONMENTS

A. G. Maher, R. W. King and J. M. Song

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

ABSTRACT - this paper examines the application of adaptive noise reduction techniques at the input to a hidden Markov model speech recognizer. The most effective technique of those discussed is spectral subtraction. As this characterizes the noise in periods of silence, it has the capability to deal with non-stationary noise sources, and is thus suitable for use in recognition systems operating over the telephone network. The paper presents results for speaker-dependent recognition of digits in gaussian white noise, and shows that the spectral subtraction noise reduction technique can maintain good recognition accuracy at signal to noise ratios as low as 5 dB.

INTRODUCTION

Telecommunications offers an important and challenging environment for applications of automatic speech recognition. Most such applications require their recognition to be speaker independent and quite reliable for a wide range of speaker environments and operating conditions.The design of robust speech recognition systems for these situations can be regarded as a problem in joint optimization of a number of processes: front-end noise reduction, acoustic parameter extraction and the speech modelling/recognition paradigm.

Recognition systems operating over the telephone face two sources of noise - that arising from the speaker's acoustic environment, and that introduced by the telecommunications network. Typically, the signal to noise ratio (SNR) at the input to the recognizer is likely to be in the range from 5 to15 dB. Furthermore, the noise statistics are quite unpredictable and may be non-stationary. These uncertainties make it impossible, in general, to train the recognizer in noise conditions identical to those of operation, as would be the optimum condition (Juang, 1991).

In this paper we examine three front-end noise reduction techniques using software simulation on a Sun IPC workstation using the ESPS/waves speech analysis environment.The speech was low pass filtered 16 bit and 8kHz (down-sampled from 32kHz) sampling rate. The weiner filtering technique, with its separate noise channel, provides a performance reference though is not suitable for implementation without the use of noise cancelling microphones in the speaker's environment. Of the two "single-ended" techniques investigated, the SNR improvement of the adaptive line enhancer (ALE) was considerably less than that obtained with the spectral subtraction (SS) method.

The overall performance of SS noise reduction at the front-end of a speaker-dependent digit hidden Markov model (HMM) recognizer is described in detail. The technique maintains good recognition performance on the standard HMM for SNR's as low as 5 dB. The paper also presents results with an improved HMM distance metric. While this somewhat overshadows the noise reduction gain for the stationary noise conditions investigated, the ability of spectral subtraction to operate adaptively and with non-stationary noise sources, validates its further study.

NOISE MODELLING AND SIGNAL TO NOISE RATIO

Most analytical work in noise reduction is based on gaussian white noise. This is mainly because it is easy to reproduce and simplifies calculations. This paper restricts discussion to white noise to simplify analysis and emphasis basic performance. The technique of spectral subtraction, examined here, is not reliant on the stationary nature of white noise.

The SNR can be computed in a number of ways. For convenience the SNR was calculated as a global measurement over the entire word as provided by the "stats" program in ESPS package.The gaussian white noise was generated using ESPS "testsd" with a random seed as generator to white noise sequence.

NOISE REDUCTION TECHNIQUES

Adaptive Weiner Filter

A Widrow-Hoff LMS FIR weiner filter (Widrow et al., 1975) was implemented as a reference. Since it requires two inputs, one being the noise reference the other the primary input of speech plus noise, it is not suitable for use in telecommunications environments without specialised equipment. If the noise reference is not contaminated with speech and the channel of noise from reference to primary input can be modelled by a FIR filter then it provides the optimum filtering.

Adaptive Line Enhancer

An adaptive line enhancer (Sambur, 1978) was implemented. This is similar to the weiner filter except the "noise" reference is derived from signal itself with a delay of the fundamental period of the speech signal and the LMS FIR section is run in reverse fashion. Instead of subtracting noise from speech as in weiner filter, the speech is subtracted from the noise. While previous work indicated it did give moderate SNR improvement, when coupled to the recognizer it failed to perform at all. This appears to be due to the amount of distortion to the signal that the ALE technique introduces. Basically the "reference" signal fails to provide the required correlation that a weiner based filter uses.

Spectral Subtraction

The spectral subtraction technique is well established (Boll, 1979), with processes shown in Figure 1. It operates by making an estimate of the spectral noise magnitude (noise bias) during periods of no speech and this estimate is subtracted from the subsequent speech spectral magnitude. The magnitude after subtraction cannot be less than zero so it can be reset to zero or some other positive floor (rectification). Boll's original algorithm tends to generate short duration narrow energy bands which give rise to unwanted "musical" tones so further modifications were required.

The spectral subtraction algorithm implemented was basically Boll's without its residual noise reduction (which failed to give improvement) but with additions suggested by Berouti et al. (1979). The major changes were the introduction of two parameters α and β. These were deigned to reduce the "musical" noise effect. The symbol α can be termed the noise bias multiplier. An overestimate of the noise bias was formed by multiplying the raw noise bias by α and this was subtracted from speech spectral magnitude. The noise bias multiplier α was in range 1 to 6. The symbol β can be termed the spectral floor multiplier. In this case the noise bias is multiplied by β and this provided the spectral floor rather than zero. The spectral floor multiplier β was in the range 0 to 0.1. For $\alpha = 1$ and $\beta = 0$ the algorithm reverts to Boll's original.

The frame size was 32 milliseconds using Hanning window and with 50% overlap between frames. The noise estimate was based on previous 5 frames of non-speech. Given the overlap this means the noise bias was based on 96ms of noise. There is an obvious trade-off in that a longer averaging will give better results for stationary noise while for semi-stationary noise the shorter the time the average is based on the quicker it can react to the changes in noise. Clearly, noise changes during continuous speech cannot be accommodated.

The most important, and as yet not resolved, part of the recognizer is the speech detection circuit. In clean speech it is relatively easy to automatically determine the speech part by simply examining energy. However as noise levels increase this simple technique fails on fricatives. A simple solution is to assume that leading and trailing frames around vowels are speech. In this case the spectral subtraction will continue to work as before but changes in the noise characteristics cannot be accommodated as quickly.

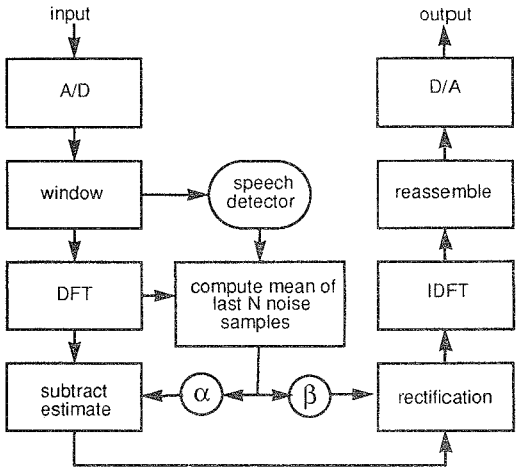

Figure 1: Block diagram of generalised spectral subtraction algorithm

Preliminary SNR results

The above techniques were tested on a short sentence and isolated phoneme pairs in various noise levels. The results are summarised in Table 1.

Input SNR (dB)	SNR improvement (dB)		
	WF	ALE	SS
5.2	20.9	5.4	10.9
9.7	16.5	3.2	6.6
19.2	7.1	-2.1	1.2

Table 1: Comparison of SNR improvement of noise reduction techniques

THE HMM RECOGNIZER

The recognizer was a 5-state HMM trained on 30 repetitions of each of the digits zero to nine. The test data was a different set of 30 repetitions of each of the ten digits to which noise was subsequently added. There were two recognizers employed in testing. While both based on the same underlying structure, the distance metrics involved are different. The first model, referred to as standard model uses the standard distance metric while the second model utilized a recent vector projection method distance metric that is more robust in noise (Carlson and Clement, 1992). Both models used 12 MFCC and 12 delta MFCC coefficients as the speech frame vectors. The models for HMM were developed using HTK (version 1.3) software.

The standard HMM recognizer (see Figure 2) shows typically good recognition performance (>95%) for SNR greater than 15 dB. The performance falls rapidly below this level. The improved model showed significant robustness to noise (around 7 dB) compared to the standard model. In fact the improved model gave the performance enhancement hoped for from the standalone noise reduction techniques.

The improved model required 20%-30% more processing time than the standard model, which is less than the time required by adding a standalone noise reduction front-end.

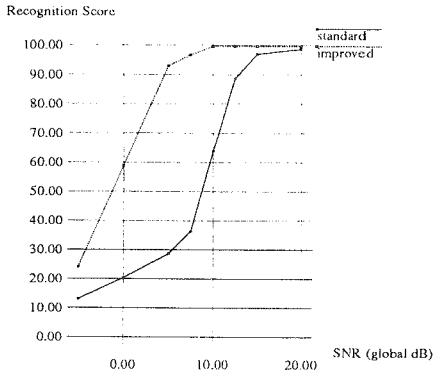

Figure 2: Performance of standard versus improved metric

TESTING PROCEDURE

The global SNR of each test utterance of the clean version of the digits was calculated and the gaussian white noise was generated and added to give the desired SNR over the range -5 to 20 dB in 5 dB increments. Then the HMM's (standard and improved) were run at each SNR consisting of 300 test samples (10 digits x 30 instances). The results are shown in Figure 2.

For testing with the spectral subtraction algorithm, the same procedure was employed but a 300ms training interval was added to the front of each digit to allow for the adaption process to work. This leading training period was removed before passing to the HMM. This was so the spectral subtraction technique itself was tested, not how well the HMM could accommodate leading silences. This technique also avoided the problem of speech detection as it was known exactly where the speech started and finished. As noted earlier, speech detection remains a problem for real implementations.

The spectral subtraction algorithm was tested with various combinations of α and β values in the ranges 0 to 6 and 0 to 0.05 respectively.

RESULTS OF NOISE REDUCTION TESTS

Figure 3 shows the result of using spectral subtraction with various values of α at a constant $\beta = 0.01$ as front-end to standard HMM. It appears that values of α around 2 to 3 give the best improvement when using spectral subtraction with the standard model. Provided β remains in range 0.01 to 0.05 there is little difference in the recognition scores. For values outside these ranges of α and β, performance degrades significantly. In the range 7 to 15 dB the spectral subtraction with standard HMM gives slightly worse results than improved HMM alone. Below 7 dB the spectral subtraction plus standard HMM shows slight improvement over improved HMM. Below 0 dB spectral subtraction gives marked improvement.

Figure 4 shows the result of using spectral subtraction with various values of α at a constant $\beta = 0.01$ as front-end to the improved HMM. In this case they provide similar results to using spectral subtraction plus the standard HMM or the improved HMM alone. Similarly provided β remains in range 0.01 to 0.05 there is little difference in the recognition scores.

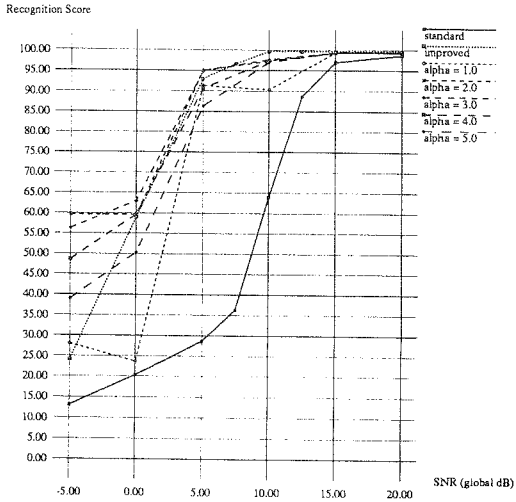

Figure 3: Varying spectral subtraction α value (β=0.01) with standard HMM

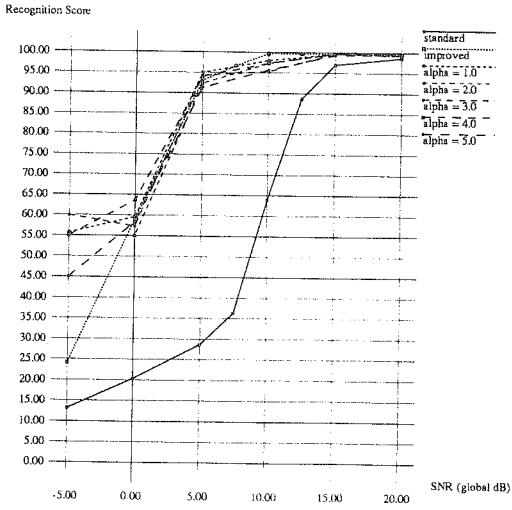

Figure 4: varying spectral subtraction α value (β=0.01) with improved HMM

CONCLUSION

The spectral subtraction technique does provide the hoped for improvement of the standard HMM recognizer but at increased computation cost. It may be possible to incorporate the spectral subtraction as part of the MFCC calculation to avoid duplication of the FFT (and eliminate inverse FFT) to reduce this cost. This remains to be investigated.

However the improved model HMM with its more robust distance metric actually gives better performance than adding the spectral subtraction front-end to the standard HMM. However these results have been generated for speaker-dependent limited vocabulary recognizer and the extrapolation to speaker independent large vocabulary recognizers may not be valid.

Also the recognizer may benefit from the spectral subtraction routine which can adapt to semi-stationary non-white gaussian noise likely to be found in operating environments.Thus the next step will to be test with more realistic noise both the standard and improved recognizers with and without a spectral subtraction front-end.The other remaining problem which requires further work is the development of reliable speech detection in noise.

ACKNOWLEDGEMENTS

This work has been carried out as part of GLASS Consortium project. The contribution and support of other GLASS partners, and the financial support of the International Division of DITAC are acknowledged.

REFERENCES

Berouti, M. et al. (1979), *Enhancement of Speech Corrupted by Acoustic Noise*, ICASSP Apr. 1979, pp. 208-211

Boll, S.F. (1979), *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans Acoust., Speech, Signal Processing Vol. ASSP-27, No. 2, Apr. 1979. pp. 113-120

Carlson, B.A. and Clement, M.A. (1992), *Speech Recognition in Noise using a Projection-based Likelihood Measure for Mixture Density HMM's*, ICASSP 92 pp. I237-I240

Juang, B.H. (1991), *Speech Recognition in Adverse Environments*, Computer Speech and Language, 5, pp.275-294

Sambur, M.R. (1978), *Adaptive Noise Cancelling for Speech Signals*, IEEE Trans Acoust., Speech, Signal Processing Vol. ASSP-26, No. 5, Oct. 1978. pp. 419-423

Song, J.M and Samouelian A. (1992), *A Robust Speaker Independent Isolated Word Recognizer over the Telephone Network Based on Modified HMM Approach*, elsewhere in these proceedings.

Widrow, et al. (1975), *Adaptive Noise Cancelling: Principles and Applications*, Proc. IEEE, Vol. 63, No. 12, Dec. 1975, pp. 1692-1716