The Institution of Engineers Australia
Communications Conference
Sydney, October 20-22 1992

A Comparison of Noise Reduction Techniques for Speech Recognition in Telecommunications Environments

A.G. MAHER, BSc, BE,
Professional Assistant,
R.W. KING, BEng, PhD, SMIREE
Associate Professor,
J.G. RATHMELL, BSc, BE, PhD, SMIREE,
Senior Lecturer,
Department of Electrical Engineering,
University of Sydney

SUMMARY The recognition performance of telecommunications applications of automatic speech recognition is likely to be degraded significantly by noise generated in the speaker's environment and by the telephone channel. This paper describes a study of techniques for noise reduction which can be applied at the input to standard recognizers trained on noise-free speech. Initial results, using signal-to-noise ratio measures, show that a spectral subtraction technique can provide up to 10 dB improvement in SNR, compared with 6 dB for an adaptive line enhancement method. Speech and pitch detection processes, which are required to support the noise reduction algorithms themselves, are also outlined.

1. INTRODUCTION

Speaker-independent automatic speech recognition has considerable potential in telecommunications environments. Applications in information and banking services have already been demonstrated (1),(2). Many telecommunications and research laboratories are engaged in fundamental and applied research to improve the scope and usability of speech recognition systems operating over the telephone. One important area of potential improvement lies in better signal processing to compensate for noise accompanying the telephone speech signal at the input to the recognition system. This noise has two components - that of the environment of the speaker, and that introduced by the telecommunications network.

It is possible to accommodate the effect of noise by training the recognizer on suitably noisy speech samples. Certainly, the training data for telecommunications applications should be collected from telephone calls. But the speaker's acoustic environmental conditions are highly variable. It is common therefore, to collect training data in controlled and relatively noise free conditions, and seek some form of noise reducing front-end signal processing to optimise the recognizer performance. This process can either be part of the recognizer front-end signal processing (as in the auditory model approach) or be an independent adaptive noise reduction process at the input of the recognizer.

Several signal processing techniques have been developed to model the way in which the human ear responds to acoustic stimuli (3),(4). These auditory models provide some evidence that they intrinsically and significantly suppress the effects of noise. They are, consequently, being integrated into speech recognition processes with some expectation of improved performance. This paper takes a more traditional approach and outlines a number of techniques which show promise for stand-alone signal pre-processing which could be applied at the input of any recognizer system. These methods rely on techniques to distinguish between speech and noise. The paper discusses these techniques and their integration in noise reduction algorithms. The paper presents some initial results of applying these noise reduction techniques to speech samples to which noise has been added. All of the work has been carried out on a SUN workstation using the ESPS Waves speech analysis software package.

2. SPEECH SOUNDS, SIGNALS, SPECTRA AND RECOGNITION TECHNIQUES

The production and characteristics of the acoustic speech signal upon which recognition systems operate is well understood and can be modelled quite accurately. The physiology of the human vocal tract imposes articulatory constraints on the range of sounds which may be generated. The sounds may be classified according to their excitation:

- voiced sounds - in which the glottis (vocal chords) modulates the air expelled from the lungs, producing a sequence of pulses separated in time by the pitch period;
- unvoiced sounds - created by turbulance of the expelled air by one of the restrictive parts of the vocal tract.

The shape of the vocal tract, the mouth and nasal cavities, can be controlled by the positions of the tongue, teeth and lips. By this means the set of sounds, or phonemes, which are used in any particular language may be defined in *articulatory* terms. To the engineer, speech production may be regarded as a slowly time-varying linear system which represents the vocal tract, driven by an excitation function which is a periodic pulse train of released air for voiced sounds and wide-band noise for unvoiced sounds. The *acoustic* signal properties of the phoneme sounds may be analysed and classified in both time and frequency domains (5). The most important analytical tool for both speech recognition and synthesis is the short-term spectrum.

Figure 1 shows the waveform and short-term frequency spectrum for the word "pan", containing the phonemes /p/a/n/. The dark areas on the spectrogram indicate the frequencies of greatest intensity. This example serves to illustrate three classes of phoneme:

- /p/ is an unvoiced plosive consonant
- /a/ is a voiced vowel
- /n/ is a nasal consonant

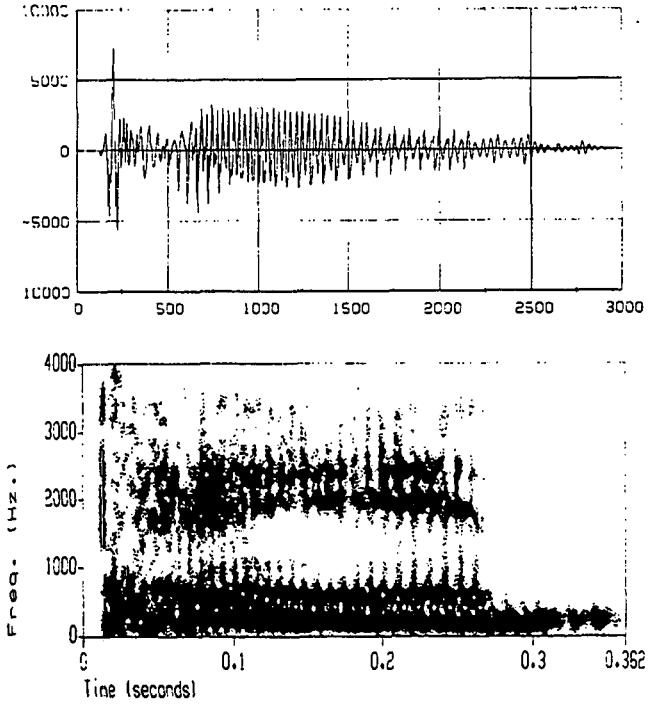

Figure 1 Speech waveform and its short term spectrogram for the word "pan"

The plosive /p/ is formed from three components. A short silence precedes the burst of sound released as the lips are opened. This is followed by an aspiration of air. The burst is a broadband sound, while the aspiration is characteristically of low amplitude. The energy in the vowel /a/ has characteristic bands of intense energy, known as *formants*. The formants correspond to the acoustic resonances in the vocal tract, and the relative frequencies of the lowest three determine the actual vowel phoneme. Like other nasal sounds, the spectrum of the /n/ has a predominant single low frequency nasal formant, in this case corresponding to the resonance of the vocal tract with the tongue blocking the mouth.

Spectrograms like that shown in Figure 1 provide a conceptual basis for identifying the requirements of automated speech recognition systems, and the problems of accompanying noise. The ability to identify the constituent sequence of phonemes is the target for large vocabulary continuous speech recognition. The natural variation in human speech makes this a daunting task even for a recognition system operating on the speaker for whom it was trained. True speaker-independent, continuous speech and large vocabulary systems require training by a large number of speakers and a significant amount of linguistic processing driven by lexical, syntactic and semantic constraints. State-of-the art systems using Hidden Markov models and artificial neural networks for phoneme recognition fall far short of the ideal levels of performance.

Speech recognition systems employ pattern matching techniques using parameters derived from the speech waveform and short-term spectrogram. To date the most successful systems have used linear prediction coefficients (LPC) of suitable all-pole filter models of the vocal tract and mel-scaled frequency cepstral coefficients (MFCC). In recognition of noisy signals the values of these coefficients will, frame by frame, be perturbed from their noise-free and ideal values. Recognition accuracy will, in turn, be reduced. Processing the noisy speech to limit the effect of noise within the recognition system is desirable. Auditory model approaches are, in effect, parameterization techniques alternative to the LPC or MFCC representation of the speech. The simpler approach, adopted in this work, is to precede any standard LPC or MFCC signal processing with an adaptive noise reduction process. The basic problem is to reduce the external noise without disturbing the unvoiced and low-intensity noise-like components of the speech signal itself.

3. NOISE MODELS AND SIGNAL-TO-NOISE RATIOS

Most analytical work in noise reduction is based on gaussian white noise. This is mainly because it is relatively easy to model and the results may be expressed in simple terms. Gaussian white noise is, however, also a reasonable model of the noise characteristics of many communication channels. In this paper we restrict the discussion to white noise in order to emphasize the basic techniques. The stationary nature of white noise forms the basis of many noise reduction techniques.

The telephone speech recognition environment is also subject to other noise and interfering sources which are non-gaussian. Ideally office noise and the interference from speakers (in the background) would be cancelled by multiple microphone techniques at the speaker's microphone. Obviously this cannot be done for all users of a telephone speech recognition system, so the task of providing general noise reduction at the input to the recognizer remains valid.

The general measure of quality of a noise reduction system is its improvement in signal-to-noise ratio (SNR improvement). This is a useful, though coarse measure, as it fails to take into account the effect of the noise reduction system on particularly sensitive sounds, such as the non-voiced sounds referred to above. The best measure for noise reduction systems for speech recognition is their improvement in recognition performance. As our aim is, at this stage, to devise techniques which can improve the performance of any recognizer trained on nominally quiet data, we shall use SNR improvement measures, and provide comments on the subjective aspects of the noise reduction process.

SNR can be computed in several ways for a nominally quiet speech signal (known as 'clean' speech) to which noise is added, depending on whether or not periods of silence are included, and on the energy balance within the speech samples. The values quoted in this paper are derived from power averages for each clean and noisy speech sample determined directly from the *stats* program provided in the ESPS Waves speech signal analysis software package.

Our experimental methodology is to add a known amount of gaussian noise to the clean speech sample, pass the noisy and clean speech through the noise reduction filter and compute the difference between the mean power of the filtered clean and filtered noise signals. This procedure eliminates problems arising with gain variations in filtering. No measures of distortion introduced by the filtering are made. The speech processing software provides visual displays of the speech signals and short-term spectrograms, and also supports audio output for subjective assessments.

4. SPEECH AND PITCH DETECTION

The noise reduction techniques described below require speech and pitch detection. Under ideal conditions (high SNR) it is easy to detect the speech by simply measuring relative energy levels, and setting a suitable threshold. Under highly noisy conditions this method becomes ineffective.

The voiced speech has relatively high energy and can be easily detected in all but the extreme SNR values. Unvoiced sounds with characteristically lower energy are less reliably detected. They are typically low in energy, often comparable to background noise with a spectrum similar to noise. Unvoiced (and voiced) speech can be discriminated from noise using zero-crossing rates. The discrimination between unvoiced speech and noise is not perfect so for safety a few frames before and after the voiced segments are considered as speech. Subjectively at least, this makes the noise-reduced speech more intelligible.

The existence of pitch is the indicator of voiced sounds used in the Sambur *adaptive line enhancer* (6) noise reduction technique described in the next section. Pitch evaluation is also important for prosodic analysis in multi-level speech recognition (7). The implementation of a robust pitch detector is itself a complex matter. Problems include unpredictable doubling or halving the actual values, although this phenomenon does not adversely affect the adaptive line enhancer.

5 NOISE REDUCTION METHODS

The principle of the noise reduction methods is to characterise the in-line noise in the silence intervals and deduct it from the subsequent speech. The methods are, essentially, adaptive filters operating on the speech spectrum. Prior to discussing these schemes we examine the ideal noise canceller from which they are derived.

5.1 Adaptive Weiner Filtering

This is a two-input technique (8) which provides a baseline for the performance of the in-line schemes. As shown in Figure 2, it has a separate input for the noise, and can, in principle, provide complete noise cancellation, through the adaptive filtering process. Clearly the scheme cannot be applied directly to the in-line noise reduction process.

The design and performance of the filter depend on the complexity and speed of the adaptation algorithm. High levels of noise reduction are typical, as described in Section 6.

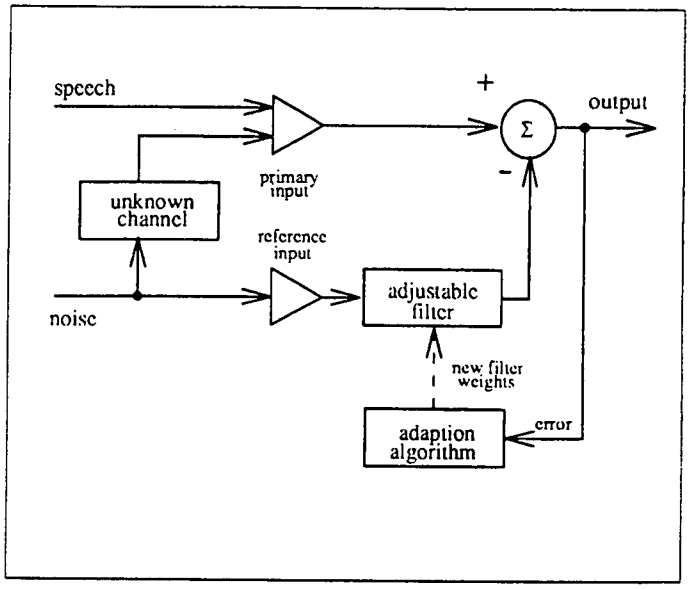

Figure 2 Adaptive Weiner filter noise cancellation

5.2 An Adaptive Line Enhancer

A block diagram of the adaptive line enhancer proposed by Sambur (6) is shown in Figure 3. This is a modification of the standard line enhancer that removes narrowband noise from a broadband signal. In this implementation the pitch detector provides a reference signal for voiced sounds, and by applying a delay of one pitch period, the noise signal can be extracted as an error signal and used to characterise the filter parameters. The output of the filter provides the speech output.

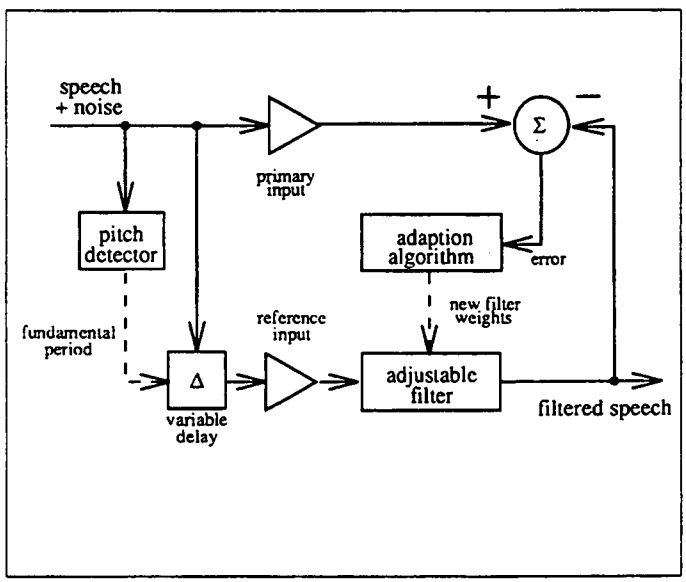

Figure 3 Adaptive line enhancer

5.3 A Multiple Microphone Technique

Like the Weiner filter, this technique cannot be applied directly at the input to a telephone speech recognition system, but it has potential for direct voice controlled computer interfaces. Furthermore it would provide a high level of performance against which the "in-line" methods like the ALE can be compared. The technique relies on the theory of beamforming which is well developed in the fields of radar, sonar and geophysics. It has also been

applied to arrays of microphones for speaker tracking in video conferencing and similar applications (9).

The technique provides noise reduction in two ways. Firstly, is the ability to (adaptively) change the direction of acoustic reception to that of the speaker, and thereby reduce the interference from other sources . The second is a standard adaptive noise-cancelling filter which characterizes the spectral characteristics of the noise and provides a direct means of adaptive filter noise reduction.

The implementation proposed by Compernolle (10), illustrated in Figure 4, requires a speech detector to decide which of the two adaptive stages should be used at any particular time. The look-direction adaptation is used when speech is detected; otherwise the noise cancelling section is applied to adapt to the noise. This system thus relies on a speech detector for good performance. The ability of the system to perform limited dereverberation is claimed to be an asset for LPC-based speech recognizers.

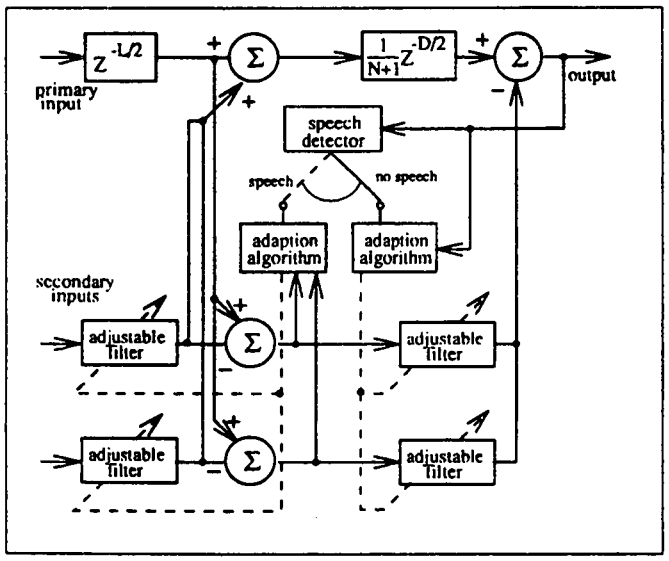

Figure 4 Three input modified beamformer for noise reduction (10)

This technique offers the best performance but at the expense of cost and applicability. It is likely to be used in a limited form in office and automobile based applications of speech recognition.

5.4 Spectral Subtraction

The spectral subtraction technique (11) is one of the most effective for our situation. It operates by making an estimate of the spectral magnitude during periods of no speech and subtracting this spectral estimate of the noise is from the subsequent speech spectral magnitude.

There are a few practical limitations to the operation of the method. Firstly, the resultant magnitude cannot fall below zero (rectification). Secondly, the technique tends to generate short duration narrow bands energy which sound like "musical" tones and which need to be eliminated. This process results in some temporal smearing of short transitory sounds, such as the burst components in plosive sounds.

In common with the other techniques spectral subtraction requires a speech detector. Because of its limitations, as noted above, leading and following frames are treated as speech. The technique is computationally expensive compared with the other techniques, due to the need to transform to and from the frequency domain, as illustrated in Figure 5. The technique is enjoying considerable attention, including the possible use of non-linear subtraction modifications have been proposed.

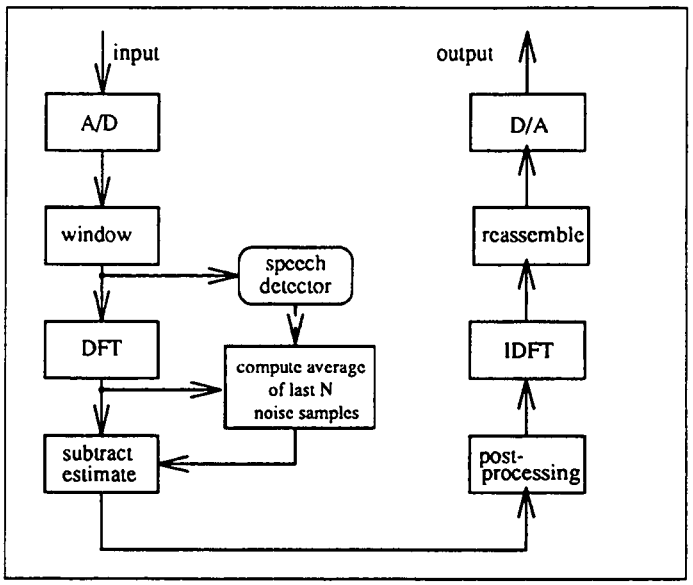

Figure 5 Spectral subtraction

6. IMPLEMENTATIONS AND COMPARISONS OF NOISE REDUCTION METHODS

The Weiner filtering (WF) and adaptive line enhancer (ALE) noise reduction algorithms have been implemented in software with a 50-tap least-mean square filtering, and tested with a number of adaptation parameters. The implementation of the spectral subtraction (SS) algorithm allows control over the speech/non-speech attenuation factor. The short-term speech spectrum is formed by a FFT routine with frames of 10 msec duration.

Two samples of clean speech to which noise was added were applied to the noise reduction processes. One speech sample is a short sentence with about 50% speech and 50% silence, the other is a set of test phoneme pairs, of which about 60% is speech. The clean speech signals used were downsampled (to 8000 samples/s) from 20 kHz sampled 16-bit linearly quantized speech. Gaussian distributed noise samples were generated for each speech sample to achieve approximately 5, 10 and 20 dB input signal to noise ratios. The noisy speech was high-pass filtered with cut-off 200 Hz to simulate the telephone channel.

Initial SNR improvement results, averaged for the two samples, are summarized in Table 1.

Clearly the adaptive Weiner filter provides the best noise reduction performance; the output signal-to-noise ratio for each speech sample approximates closely to the maximum possible value for that sample. The in-line processes do not perform as well, although the spectral

subtraction method provides significant improvement for the lower input SNR values. As noted earlier, this method is somewhat more complex than the adaptive line enhancer. The degradation found for the ALE operating at high input SNR is, at present, unexplained.

TABLE 1.

COMPARISON OF PERFORMANCE OF NOISE REDUCTION TECHNIQUES

SNR input	SNR improvement, dB		
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

Subjectively too the spectral subtraction technique performs very well on reducing noise in the intervals of voiced speech. In its present implementation, it tends to cut off rather too severely the unvoiced components of speech. This will be examined in detail in future work and in combination with a Hidden Markov model phoneme recognizer.

7. CONCLUSION

The results described above are our initial findings in this study of noise reduction techniques. The spectral subtraction method clearly has significant potential merit. The speech and pitch detection methods, adaptation parameters, filter lengths deserve further attention. The eventual merit of these techniques will be found in their integration within speech recognition systems. In this regard we plan to examine their performance for phoneme recognition performance in HMM recognizers with standard and in comparison with auditory model front-ends.

8. ACKNOWLEDGEMENTS

This work has been carried out as part of the GLASS (Generalised Language and Speech System) Consortium project funded by the International Division of the Australian Department of Trade, Industry and Commerce. Their financial support is gratefully acknowledged.

9. REFERENCES

1. Wheddon, C. "Human-Computer Speech Communication Systems", Proc. 2nd Aust. Conf. on Speech Science and Technology SST-88, Sydney, Nov. 1988, pp. 256-261.

2. Wilpon, J.G., Mikkilineni, R.P., Roe, D.B. and Gokcen, S. "Speech Recognition: from the Laboratory to the Real World", AT&T Tech. Jnl., Vol. 69, No. 5, Sept/Oct 1990, pp 14-23.

3. Seneff, S. "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", Jnl. of Phonetics, Vol. 16, 1988, pp 55-76.

4. Ghitza, O. "Auditory Nerve Representation as a Front-end for Speech Recognition in Noisy Environments", Computer Speech and Language, Vol 1, 1986, pp 109-130.

5. O'Shaunessy, D. "Speech Communication: Human and Machine", Addison Wesley, 1987

6. Sambur, M.R. "Adaptive Noise Canceling for Speech Signals", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-26, Oct. 1978, pp. 419-423.

7. Rowles, C., Huang, X., and Aumann, G. "Natural Language Understanding and Speech Recognition: Exploring the Connections, Proc. 3rd Aust. Conf. on Speech Science and Technology SST-90, Melbourne, Nov. 1990, pp. 374-379.

8. Widrow, et al. "Adaptive Noise Cancelling: Principles and Applications", Proc. IEEE, Vol. 63, Dec. 1975, pp. 1692-1716.

9. Flanagan, J.L., Johnstone, J.D., Zahn, R., and Elko, G. "Computer-steered Microphone Arrays for Transduction in Large Rooms", Jnl. Ac. Soc. Am., Vol. 78, Nov. 1985, pp 1508-1518.

10. Van Compernolle, D. "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings" IEEE Int. Conf. Acoust. Speech & Signal Processing, Albuquerque, Apr. 1990, pp. 833-836.

11. Boll, S.F. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, Apr. 1979, pp. 113-120.