USE OF SPEECH RECOGNITION SYSTEMS IN NOISY ENVIRONMENTS

A.L. Harvey, H.J.N. Lee

Department of Electrical Engineering
Royal Melbourne Institute of Technology

ABSTRACT - This paper discusses the use of discrete utterance speaker dependent recognisers in acoustically noisy industrial environments. Aspects discussed are training of the recogniser system, spectral subtraction, type of features to be extracted and the effect of vocabulary size. The author worked with Intel's speech recognition group for six months during the early stages of a voice data entry contract with General Motors.

INTRODUCTION

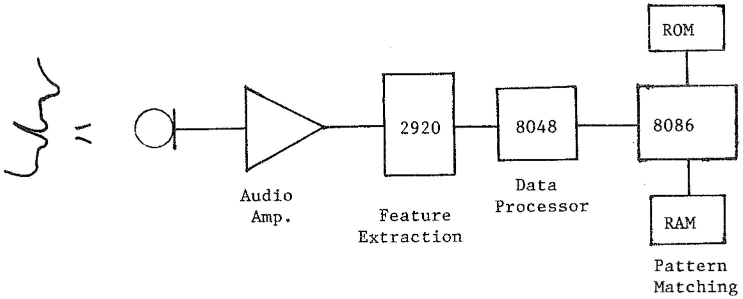
Information entry to computer data bases by means of speech utterances has been used commercially for some years. The use of recognisers using linear predictive coding or filter banks for feature extraction is quite feasible in relatively quiet office environments. However generation of inspection data in factory environments where a high noise level is normal adds problems to an already fragile technology. The techniques to be discussed were investigated by Intel in applying their recogniser to a vehicle inspection application at General Motors in Detroit.

G.M. INSPECTION DATA VOICE INPUT INVESTIGATION

G.M. had tried out several recognisers for their shop floor vehicle inspection data base application but found that LPC type recognisers were not as reliable as filter bank recognisers in higher noise environments. On this basis Intel were given the contract for finished vehicle inspection information data base generation.

INTEL'S SPEECH RECOGNITION SYSTEM

A block diagram of Intel's system is shown below. The system has a maximum vocabulary of 200 discrete utterances and is speaker dependent.



Amplified speech, band limited to 3.3 kHz was sampled at 8.0 kHz by the 2920. The speech energy was then applied to sixteen band pass filters approximately 200 Hz wide. The rectified output of the filter bank was then sampled every 8.0 milliseconds by the 8048 and the spectral slope between adjacent energy bands quantified as +1 for positive slopes, 0 for equal energy peaks (flat slope) and -1 for negative slopes.

The data was then time sliced into twelve groups and then each group totalised.

Input energy level was manually adjustable by means of an audio gain control. Level indicating lights, driven by software reading average level were used to set the gain control. Recognition time was about 10 msec/word.

UTTERANCE END POINT DETECTION

Initially a data frame above an adjustable threshold energy level was used to trigger the data input sequence. To reduce triggering of the system by odd high intensity noises, e.g. hammer blows, a system was introduced in which two or more consecutive high energy frames were required to trigger the recogniser system.

Longer noise pulses, while triggering the recogniser would always fail safe as the recognition score would always be very low. Operation under near zero signal to noise ratios is not possible with the above system. However using a pushbutton to generate word end points did give usuable operation. Recognition scores around 95% wave obtainable on vocabularies around 50 words.

TRAINING PROCEDURE

The training of utterances, (words or word groups) was investigated by the speech team. Training under the same conditions as recognition gave the best results. Therefore for noisy environments, training in noise was required. Recognition errors increased as the signal to noise ratio went down but not nearly as rapidly as when training in silence and recognising in noise.

All utterances were trained three times and an average taken weighted in favour of the most recent utterance.

ON-LINE TRAINING

A speech pattern would be updated using the most recent pattern automatically as long as the recognition score was high, above about 80%.

If the utterance was consistently rejected, a retraining procedure could be invoked either by keyboard commands or verbally, using the word retrain.

SPECTRAL SUBTRACTION

As discussed earlier, training in noise gave better results over training in silence for use of the voice recognition system in the same noise. However recognition scores did deteriorate, as expected as the noise level increased. The reduction in orthogonality (spectral uniqueness) of the (S+N) spectrum reduced the difference in recognition scores compared to noise free training and recognition.

Spectral subtraction techniques were used with better results than using noisy templates and noisy recognition environments.

In this method, the noise spectrum was obtained during normal factory operation by triggering the recogniser manually. The resulting noise spectrum was then subtracted from both templates and recognition utterances. This technique (S+N)-N=S improved the uniqueness of templates but only if the noise spectrum was reasonably stationary and equal to the noise spectrum during the recognition phase.

USE OF SPEECH FEEDBACK

The addition of a synthesiser board to give speech feedback to the operator closes the loop on the speech recognition system. This can be done visually of course using a VDU but this means the operator must stop the inspection process to look at the screen output. Verbalising the recognised word gives a rapid answer to the operator regarding which word was recognised.

No commercial system at the present state of the art of speech recognition has attained 100% recognition accuracy consistently.

The Intel system quotes 99% accuracy for its 50 American city vocabulary. Therefore error checks are necessary.

VOCABULARY SELECTION

Selection of spectrally unique utterances is one of the most important phases of a successful speech input system. Words causing consistent substitution errors must be deleted and a word with a similar meaning substituted.

The speech synthesiser is also very useful here as it may be programmed to provide prompting of the operator so that he does not need to remember what the chosen vocabulary words were. For larger vocabularies, synthesiser feedback becomes a very important part of the system, both for error checking, vocabulary prompting and a systematic execution of the inspection process.

VOCABULARY SIZE

In any speech recognition system it is desirable to keep vocabulary size to a minimum.

For large vocabularies the vocabulary will have to be broken up into several smaller vocabularies to give acceptable recognition accuracy. This technique is even more important in noisy environments. Smaller word groups (clusters) may be necessary with the reduced system discrimination in acoustically noisy environments.

PERFORMANCE OF AN LPC RECOGNISER IN NOISE

An eight stage linear predictive coding (LPC) based speech recognition system was developed and tested in the presence of acoustic noise. The system extracted eight reflection coefficients and two data reduction algorithms were developed to investigate the effect of data reduction methods on recognition accuracy.

The following data reduction methods were used on the LPC data using

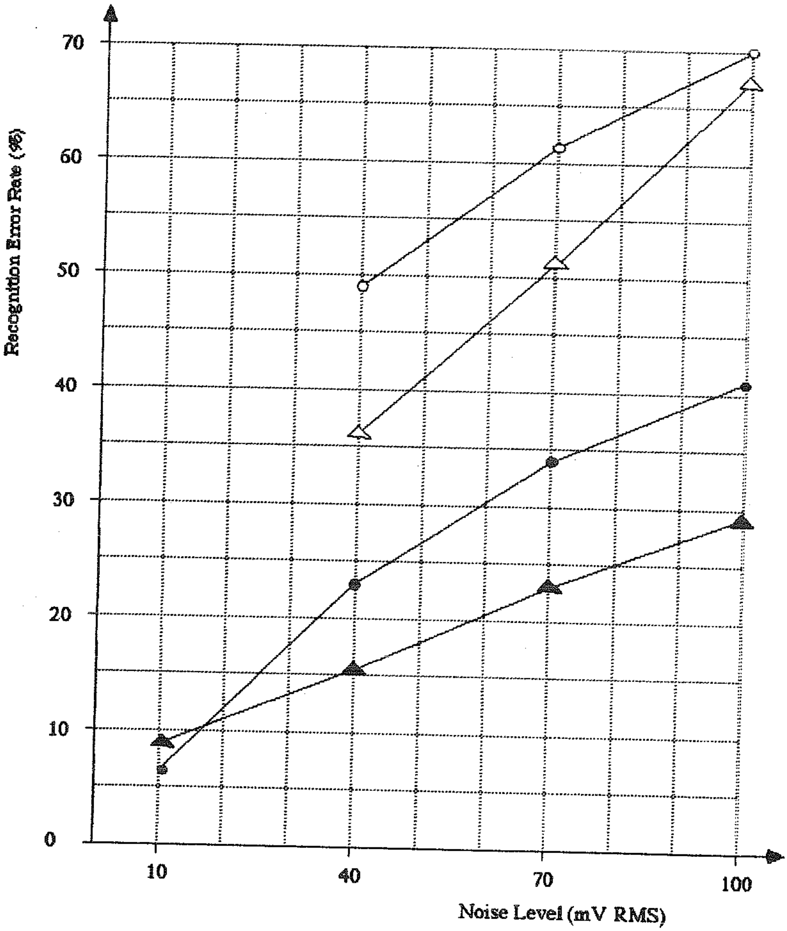


Figure 1. **Effects of Noise on the System Performance.**

▲ DC-DMCC method
● APTN-Warp method] Trained in noise conditions.

△ DC-DMCC method
○ APTN-Warp method] Trained in quiet condition.

(i) Direct compression (DC) (Lee, Bradley and Harvey, 1986)
(ii) Amplitude preservation and time normalisation.

Direct compression is a method of data reduction in which the LPC data for an utterance is split into two sets of frames and a modified average of each LPC coefficient is taken for each set of data.

Amplitude preservation is a method of data normalisation in which the utterance LPC data is normalised to a standard length and a weighting is put on the coefficients to preserve the area of the utterance.

Figure one illustrates the increased recognition error rates when training in silence and recognising in noise compared to training in noise and recognition in the same noise.

Very approximate signal to noise ratios may be obtained by using a speech signal level of 0.5 volt RMS.

CONCLUSIONS

Single utterance, speaker dependent recognisers at least of the filter bank type can be used in acoustically noisy industrial locations.

Training should be done under the same conditions as recognition. This will be particularly effective for reasonably stationary noise spectrums. Spectral subtraction will give better results but is more complicated to perform.

Word end detection algorithms must give reasonable performance in noise, normally meaning a high threshold level maintained for several frames of data.

Vocabularies should be kept small by breaking up larger vocabularies into subsets on some logical basis.

REFERENCES

LEE, H.J.N., BRADLEY, A.B.B. & HARVEY, A.L.(1986) "Isolated Word Recognition based on Direct Compression". IEE Conference on Speech Input/Output. London.

REDDY, O.R.(1976) "Speech Recognition by Machine : A Review". Proc.IEEE, Vol.64. April 1976.

ROLLINS, A. & WIESON, J.,(1983) "Speech Recognition and Noise". IEEE ICASSP. PP 523-526, 1983.

HOY, L., BURNS, B., SOLDAN, D. & YARLAGADDA, R. (1983) "Noise Suppression Methods for Speech Applications". IEEE ICASSP, pp 1183-1136, 1983