[From: *Translation and Communication: Translating and the Computer 6*. Proceedings of a conference ... 20-21 November 1984, ed. Catriona Picken (London: Aslib, 1985)]

Automatic speech recognition: fact or fantasy?

Raj Gunawardana

Speech Systems Manager, Bedford Regional Technology Centre, Texas Instruments Ltd, Bedford, UK

INTRODUCTION

The application of speech as a communication medium for reporting machine status conditions, alarm alert etc. has been a technological reality for a number of years. However, social awareness of speech recognition has been confined mainly to science fiction portrayals, such as the intelligent on-board computer 'HAL' in the film 2001 - a space odyssey. Although the detail of the human-like intelligence demonstrated by 'HAL' still remains exclusive to science fiction, some of the 'functions' of voice communication have already been shown to be achievable by the use of modern VLSI (Very Large Scale Integration) technology. This paper sets out to discuss the particular problems associated with automatic speech recognition and the current state of the art.

THE TASK

An automatic speech recognition system will be expected to produce a particular machine response to a voiced stimulus. As shown in Figure 1, the task would be one of detecting voice and translating a voice signal to a form that can be handled by the machine, then identifying an utterance as a unique one and effecting the response associated with the 'Voiced Event'. Assuming that there will be performance restrictions, a machine that can discriminate between the words 'yes' and 'no' will, in effect, appear at a first glance to be a simple recognition system. However, even at this

level of performance, the limitations will not be simple. Much like human beings, the machine will make errors. For example, the machine can 'substitute' one of its responses for either another input listed within its capability or one that is outside its 'recorded list'. Equally, an input which is listed may be 'rejected' by the machine. The extent to which such errors occur may depend on the speaker making the voiced input. It follows, therefore, that a number of performance criteria can be established to 'measure' and hence describe the features of a speech recognition system. Naturally the levels of performance that can be achieved are related to the cost of implementation.

This paper does not attempt to discuss either the particular problems associated with translation of language or any possible solutions. Instead it will address the following topics:

a. Speech signal and perceived intelligence.
b. Processing the speech signal for recognition.
c. The market for speech recognition.
d. Cost and performance of products available.
e. Conclusions.

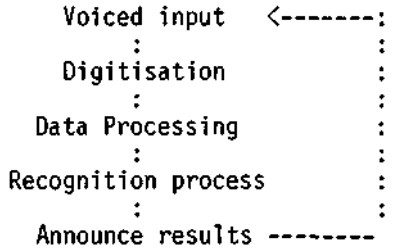


Figure 1. Automatic speech recognition

THE SPEECH SIGNAL

The complexity of speech processing is related to the vast amount of 'acoustic data' present in the speech waveform. However, the problem is not so much handling the data as extracting only the spoken intelligence within the data. It can be concluded that data associated with 'perceived intelligence' will be low because of the limitations of human nerve capability in conveying these data. The human speech generating mechanism restricts the voice signal bandwidth to around 4kHz. This means that in converting such a signal into digital form, a data rate of 96,000 bits will be generated, allowing for a sampling rate of 8 kHz to capture the full bandwidth and a sample resolution of 12 bits to encompass a dynamic range of about 65 dBs. However, it is possible, by a non-linear quantisation of the signal, to reduce the sample resolution to 8 bits without perceptible loss of quality. This technique is commonly adopted in telephone communications in Log-Companded PCM, producing a reduced data rate of 64,000 bits per second. Even at this rate of 8 Kbytes of data every second, it will not be economical for a low-to medium-cost computer to store any reasonable amount of data or to cope with speeds associated in dealing with such a quantity of data.

The speech signal (waveform) contains much that is redundant, as far as human perception is concerned. For example, it is established that perceived intelligence is contained in the dynamics of a number of spectral peaks (known as formants) that occur in a speech signal. Consequently, by processing the speech signal to derive this information and thereby losing all time-variant waveform information contained in the signal, it is possible to reduce the speech data rate/bandwidth to a considerable extent. Linear Predictive Coding (or LPC, as it is known) permits the data rate to be reduced to as little as 1,000 bits per second, equivalent to about 125 bytes per second. LPC does this by effectively retaining a minimum of information to model the voice spectrum, and subsequently modelling the human vocal tract by assimilating the vocal tract into a digital filter and using one of two excitation sources to implement the basic sound generation mechanisms (Figure 2). LPC has been used frequently for voice coding and synthesis applications due to its low data rate. It is equally useful for speech recognition applications.

There are, of course, other techniques of processing speech data to reduce its data rate, but these are often too inefficient or too costly to implement.

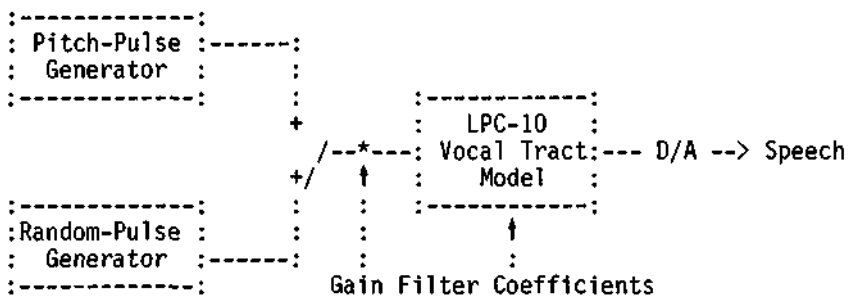


Figure 2. Vocal tract model

SPEECH SIGNAL PROCESSING FOR RECOGNITION

Most recognition systems currently available perform the task by pattern matching. That is, incoming speech data are compared with data stored within a recognition system, and a best-fit pattern matched. The stored data in effect represent all or a part of the 'perception intelligence' features of speech. The same characteristic features have to be extracted from incoming data to make a coherent match.

As discussed previously, the intelligence features are contained with the spectral peaks known as 'formants'. A number of alternative techniques which can be used to code these to a reasonable data rate for fast processing are discussed later. Once the features have been extracted, it is necessary to ensure that the time relationship of the incoming data is correctly aligned. Once this is done, a measure of similarity could be made with all the corresponding patterns stored within the reference data. The selection of the best possible match is itself a substantial task when there are a number of possible candidates identified. Once the best match is identified, the final task of announcing the results will be simple, depending on the type of application that the recognition system is being used for.

Feature extraction

One of the simplest means of extracting speech recognition data is to count the changes in the polarity of the incoming speech signal over a fixed period (or 'window'). This technique (the 'zero crossing technique') provides an

indication of the value of the predominant frequency of a speech signal and, hence, can be used only for discriminating an absolute minimum number of words. The performance of a system incorporating such an approach will also be very low. An alternative approach is to make a direct measurement of the spectral energy distribution of speech. This can be achieved by distributing a bank of filters evenly over the voice band and exciting each filter with a speech signal. The response of each filter will indicate the presence and magnitude of energy associated with the relevant frequency. The disadvantages of this approach are that it is costly to implement and produces inadequate resolution of frequency, especially where the formants are located very close to each other. One method that has gained in popularity is to model the speech signal using LPC parameters. This has the advantage of retaining all the detail associated with the voice spectral envelope. However, it also calls for considerable computing capability and only with the advent of high speed signal micro-processors using VLSI technology have the costs of this method become reasonable.

Time registration

A simple way to match time alignment of an incoming speech signal and a template is to determine the endpoints of the spoken data and then to compress it or stretch in time accordingly. However, this has a number of disadvantages. The technique imposes isolation of each word (utterance) from any following utterances. The recogniser will have to confine itself to simply extracting features of speech, since measuring similarity will not be possible until the utterance length is known. Any non-linearity in alignment of data, as is frequently present in speech, will result in poor similarity measure due to time rather than spectral mismatch. A technique that overcomes this is dynamic time warping (DTW). This technique optimises the 'distance' measure between reference and speech features by adjusting time registration non-linearly. The task can be carried out in synchronism with speech and hence allows recognition of continuous speech.

Decision strategy

Similarity measurement in pattern matching often yields results where a multiplicity of patterns become potential candidates for a result. This can be overcome by statistically evaluating the results obtained. One way to enhance the performance is to exploit the finite states associated with a given vocabulary. This works well with a

small vocabulary but becomes less practical as both the vocabulary and user freedom increase. Equally, the problem will not be solved for isolated utterances. Currently, research into speech recognition is being focussed on detecting statistically weighted phonetic sequences which will not only enhance the performance of substitution and rejection, but will possibly permit larger vocabulary and speaker-independent performance.

IMPLEMENTATION IN HARDWARE

A low cost speech recognition peripheral can be implemented using the Texas Instruments TMS32010 signal processor, as shown in Figure 3. This architecture will be able to effect many other speech processing functions as well as recognition (e.g. verification, coding etc.).

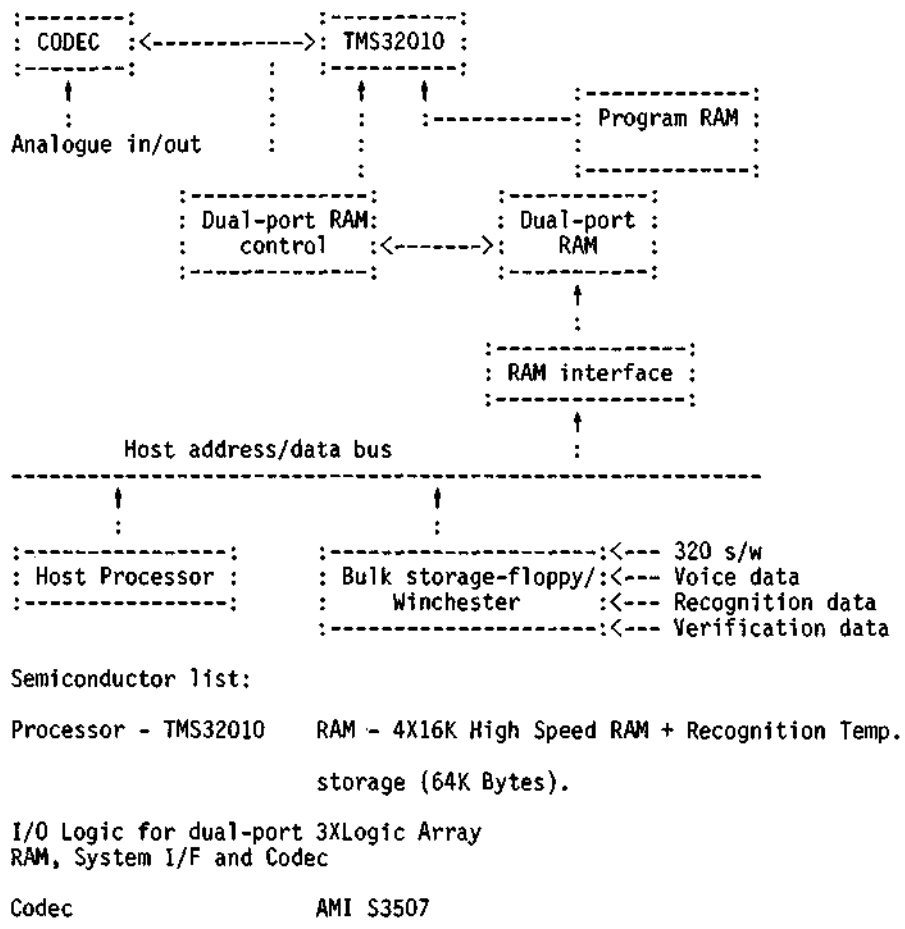
SPEECH RECOGNITION

Speech recognition systems/algorithms currently available fall into two main categories. These are:

- Speaker dependent (SD)
- Speaker independent (SI)

Speaker dependent recognition

The speaker dependent systems require the machine to be trained to the particular speaker's voice characteristics and will, in the 'training mode', create a number of 'templates' for subsequent use in recognition. During the training mode, the user has complete freedom to create the vocabulary required (within the limits of its extent). During the 'recognition mode', the system will analyse an utterance and will attempt to match the data to one of the templates previously created. In other words, the system will make a measurement of similarity, or 'distance' (as it is often called). In measuring similarity, it is necessary to ensure that the points being compared with the stored reference match the incoming data in time as well as pattern. This process is called 'time registration' alignment and can be done by end-point determination (of the utterance) or by 'warping' time non-linearly, using an algorithm known as 'dynamic time warping'. Having identified the most similar measurement, the 'candidate's' similarity is compared with a pre-determined threshold and its acceptance or rejection is announced.



Semiconductor list:

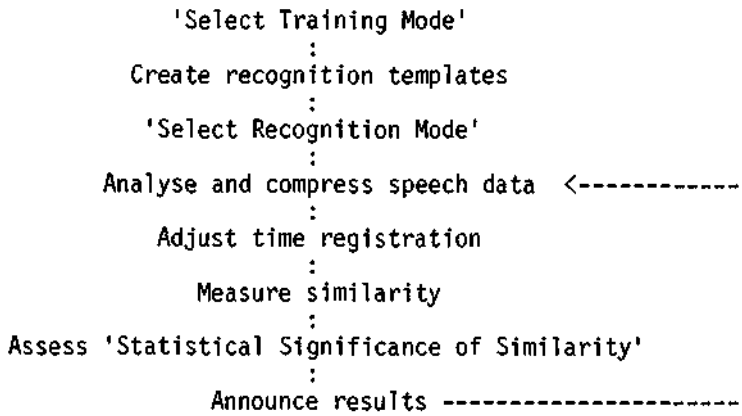
Processor - TMS32010	RAM - 4X16K High Speed RAM + Recognition Temp. storage (64K Bytes).
I/O Logic for dual-port RAM, System I/F and Codec	3XLogic Array
Codec	AMI S3507

Figure 3. Low-cost speech recognition using the TMS 32010 signal process

Two measures of speech recognition are frequently used to describe system performance:

1. Rejection error occurs when the recogniser 'rejects' a word known to be within the vocabulary. Such errors, although annoying to the user, do not bring about serious system errors (erasure of a file).
2. Substitution error is when a different word is 'substituted' for an utterance within the vocabulary, outside it, or simply environmental noise. Such errors can obviously produce drastic results. Good to medium performance recognisers have rejection error rates of the order of 1 per cent, and a substitution rate of about half this figure.

The process of speaker dependent recognition will therefore be that shown in Figure 4.



Speaker independent recognition

In speaker independent (SI) recognition systems, both functionality and performance differ from SD recognisers. Reference data/templates are created as part of a prototyping process and are in effect fixed for ever. The creation of such data is often a major linguistic exercise involving analysis of properties of data from a large population of speakers. Having created data that are a 'median' of speech, it is possible to get recognition of speech independent of speaker. However, the extent of 'substitution' and 'rejection' error rates are significantly higher than for speaker dependent systems.

Nevertheless, the current challenge in speech technology appears to be to produce high performance SI algorithms. The trend appears to be to incorporate language-related

processing into the algorithm to improve performance. Amongst the approaches being considered are techniques exploiting 'finite states of grammar' and 'phonetic segmentation'/matching algorithms. The latter in particular is advantageous in that once a phonetic database is established for a given language, it will be possible for a linguist to define the sequence to be expected in a new word added to a vocabulary. Implementation of such techniques will be made possible by using signal processors such as the TMS320. These algorithms are likely to be available in late 1985.

THE STATE OF THE ART

There have been speech recognition systems operating in mini- and mainframe computer environments for some time. Needless to say, these systems have been very high in cost and have been used only in very specialist applications where speech recognition requirement has been essential and where cost has been of relatively low consequence. These systems have had typical features which are ideally required in speech recognition, systems such as multi-speaker capability, access to other information, incorporation of language into the performance, etc. Such systems are still being perfected and it can be expected that the cost will be tolerated in return for high performance. However, the most volatile area of speech recognition systems is in the product range that can be described as being medium in cost. These systems have come about as a result of the availability of high-speed signal processors. Characteristically, they have simply borrowed the algorithms used in the traditional minicomputer-based recognition systems. Hence, when it comes to actual performance, they compare very favourably with their more expensive counterparts. However, due to the inevitable restrictions that apply to a signal processing device, these systems are found effectively 'bolted' to a host computer (typically a low-cost 'personal computer'). One such popular computer has some dozen or so manufacturers producing add-on modules offering speech recognition capability. Medium-cost systems currently available continue to have restrictions of active vocabulary size, typically being 50 words or so, and are likely to be speaker dependent recognisers.

There are some speech recognisers in the market which achieve low cost levels by using dedicated hardware to perform recognition. However, to date they still have severe limits of vocabulary size and performance. Nevertheless, it can be expected that with improvement of VLSI capability, there will be considerable progress in

products falling into the low cost category. Many applications can tolerate or overcome the performance limits that may prevail in low cost recognisers. The problem to overcome will mainly be one of coping with 'customisation' of a design, to meet the particular language requirements as well as vocabulary.

THE MARKET

There is ample evidence that there is considerable market pressure for products with speech recognition capability. One would expect to find useful applications in a whole range of industries including communications, computers, consumer goods, and so on. However, it has to be admitted that there is a mismatch between what today's technology is capable of offering and the expectations of the marketplace. These expectations cover capabilities such as speaker independence, connected speech recognition, large vocabulary capacity, and high performance in rejection and substitution errors. Additionally, the type of applications which offer substantial volume markets are also very intolerant of the cost requirements at the current time. Nevertheless, it would be fair to assume that, given the opportunity for commercial capitalisation, the challenge will be picked up by the technologists to provide the necessary solutions to achieve success.

CONCLUSIONS

Given the current focus of interest of technology and given the current trends, it is likely that there will be applications related predominantly to business terminals, particularly those linked to networks to provide communications capability. It is likely that machines capable of human-like intelligence, so-called 'artificial intelligence', will lead the way to overcome some of the limitations that are proving to be a bottleneck for a profusion of applications. The timing of any major breakthrough is much more difficult to forecast. It may that there will be no revolutionary change. Instead, we may see an evolutionary change developing fairly rapidly with the increase in complexity resulting from VLSI technology. When a machine will be able to understand humans as we understand each other is an arguable point but that it will come to pass in time is an inevitability.

AUTHOR

Raj Gunawardana, Speech Systems Manager,
Bedford Regional Technology Centre, Texas Instruments Ltd,
Manton Lane, Bedford MK41 7PA, UK.