# A MICROCOMPUTER BASED SPEECH RECOGNITION SYSTEM

Edward J. Webb, K. Menon, and Ching Y. Suen
*Department of Computer Science*
*Concordia University*

## ABSTRACT

This paper describes the properties of speech and a system which recognizes speech by means of a microcomputer system. It is implemented on an 8-bit microcomputer which uses a Z80 central processing unit and an analog to digital interface for the acquisition and preprocessing of the speech signal obtained from a microphone. The system is designed to adapt itself to recognize words from different speakers by changing the stored acoustic parameters. Experiments on a large number of samples show that after careful training, recognition of isolated words is attained with a high degree of accuracy.

# UN SYSTÈME POUR LA RECONNAISSANCE DE LA PAROLE, BASÉ SUR UN MICROPROCESSEUR

## RÉSUMÉ

La présente communication décrit les propriétés de la parole et un système de reconnaissance de la parole utilisant un micro-ordinateur. Ce système est mis en oeuvre dans un micro-ordinateur à 8 bits dont l'unité centrale est une pastille Z80, et qui comprend un interface analogique-numérique pour la saisie et le pré-traitement du signal audio capté par un microphone. Il a été conçu pour s'adapter de lui-même à la reconnaissance des mots prononcés par diverses personnes, en modifiant les paramètres acoustiques mis en mémoire. Des expériences portant sur un grand nombre d'échantillons montrent, qu'à la suite d'un apprentissage minutieux, ce système peut reconnaître avec un degré élevé de précision des mots détachés.

## INTRODUCTION

The desirability as well as the need for people to communicate with machines in the natural mode i.e., the human voice, has been a major impetus to the growth of man-machine communication by voice. The ability to provide voice communication would allow complex machines to be accessed and controlled by large groups of people with little special training. Another reason for this widespread and rapid growth in the area of man-machine communication by voice has been the increase in capability and the decrease in cost of modern digital hardware. The development of speech communication systems can have an important economic impact due to their unique ability to communicate data or commands to and from an operator in environments involving access control for security purposes, validation of the speaker over the telephone for purposes such as banking transactions, credit purchases, use of telephone credit cards, etc., or in environments involving inventory control, automatic material sorting, quality control stations in manufacturing plants.

In general, the field of man-machine communication by voice can be categorized into three broad areas: 1) voice response systems, 2) speaker-recognition systems, and 3) speech-recognition systems.

While voice-response systems communicate by voice in one direction only - from the machine to the user - speaker recognition and speech recognition systems communicate *from* the user *to* the machine. A speaker recognition system can be divided into two areas, i.e. speaker verification (to decide whether the speaker is one whom he claims to be), and speaker identification (to determine the speaker's identity from a prearranged list of speakers).

The basic task of speech-recognition systems is either to recognize the entire spoken utterance exactly, or else to "understand" the spoken utterance. The concept of understanding rather than recognizing the utterance is important for systems which deal with continuous speech input incorporating a fairly large vocabulary. On the other hand, exact recognition is of primary importance for systems which deal with limited vocabulary, isolated-word, and restricted number of speakers [1].

A native speaker uses his knowledge of the language, the environment, and the context, in understanding a spoken sentence. This knowledge may be classified as follows: phonetics, phonology, prosodics, lexicon, syntax, semantics, and pragmatics; and they cover increasingly larger units of speech [2]. In other words, in a speech recognition system, the knowledge sources at a particular level must be made to communicate and cooperate with those at another level, in the process of decoding and utterance. Steps must also be taken to prevent errors at one level from corrupting the knowledge sources at other levels. However, even a perfect phonetic transcription of a spoken sentence can make the detection of word boundaries ambiguous, if not difficult, owing to the phonological influence across adjacent words and inconsistencies in enunciation [3]. One strategy would be, to adopt a system in which there is minimal communication among

knowledge sources, e.g., an isolated-word speech recognition system with limited vocabulary. In such a system, the individual words are spoken with a pause (100 ms minimum) between words. Such pauses will allow us to establish the word boundaries with less ambiguity, resulting in high recognition accuracies [4]. This paper describes such a system implemented on an 8-bit microcomputer.

## SYSTEM HARDWARE

Our speech recognition system is implemented on an 8-bit microcomputer which uses a Z80 central processing unit. The vocal input is given at the microphone which is also used as a beeper when driven by the signal generator (fig. 1). A beep from the microphone provides a cue to the speaker when the signal generator is enabled. This beep, of 10 ms duration, signals the beginning of the speech window and a second similar beep 1.5 sec later signals its end. The speech input is differentially amplified by the microphone preamplifier up to 5kHz to compensate for the greater energy found in the lower frequency components of speech. This pre-emphasis tends to balance the three formant values collected in the next stage and compensate for the attenuation of higher frequencies in the vocal tract.

Three band-pass filters receive the output of the preamplifier and extract the portion of the speech signal which falls within the frequency range of the filter. The filter values are chosen to approximate the frequencies of formants 1, 2 and 3 of the average vocal tract. The range of filter F1 is 150Hz to 900Hz, F2 is 900Hz to 2200Hz and F3 is 2200Hz to 5000Hz. Each filter output is a sinewave proportional to the microphone input within the filter passband and varies about a level of 2 volts. A time averager (TA1, TA2 and TA3) follows each filter, detecting the peak of the input waveform and averages this over a fixed time period generating a voltage from 0 to 4 volts which is proportional to the signal within each band.

The amplified microphone output from A1 is also applied to a second amplifier A2 where it is again emphasized up to 5kHz and its unfiltered waveform swings $\pm$ 2 volts about a 2-volt rest level and is passed directly to the analog multiplexer M1. The output from A2 also passes to a zero-crossing detector (ZCD) which produces a voltage proportional to the number of times the amplified speech signal crosses its rest level in a given time and thus serves as an approximate measure of the frequency of the spoken input.

The outputs from the time averagers, the zero-crossing detector and three reference voltages which are used for testing the speech system are available at the 8 to 1 analog multiplexer. The selected output is passed either directly to the analog to digital converter or it is routed through a compression amplifier by a 2 to 1 analog multiplexer M2. The compression amplifier increases the low amplitude signals and reduces those of higher levels approximating a logarithmic gain of its output signal. This affords a means of compensating for the variation in the volume of

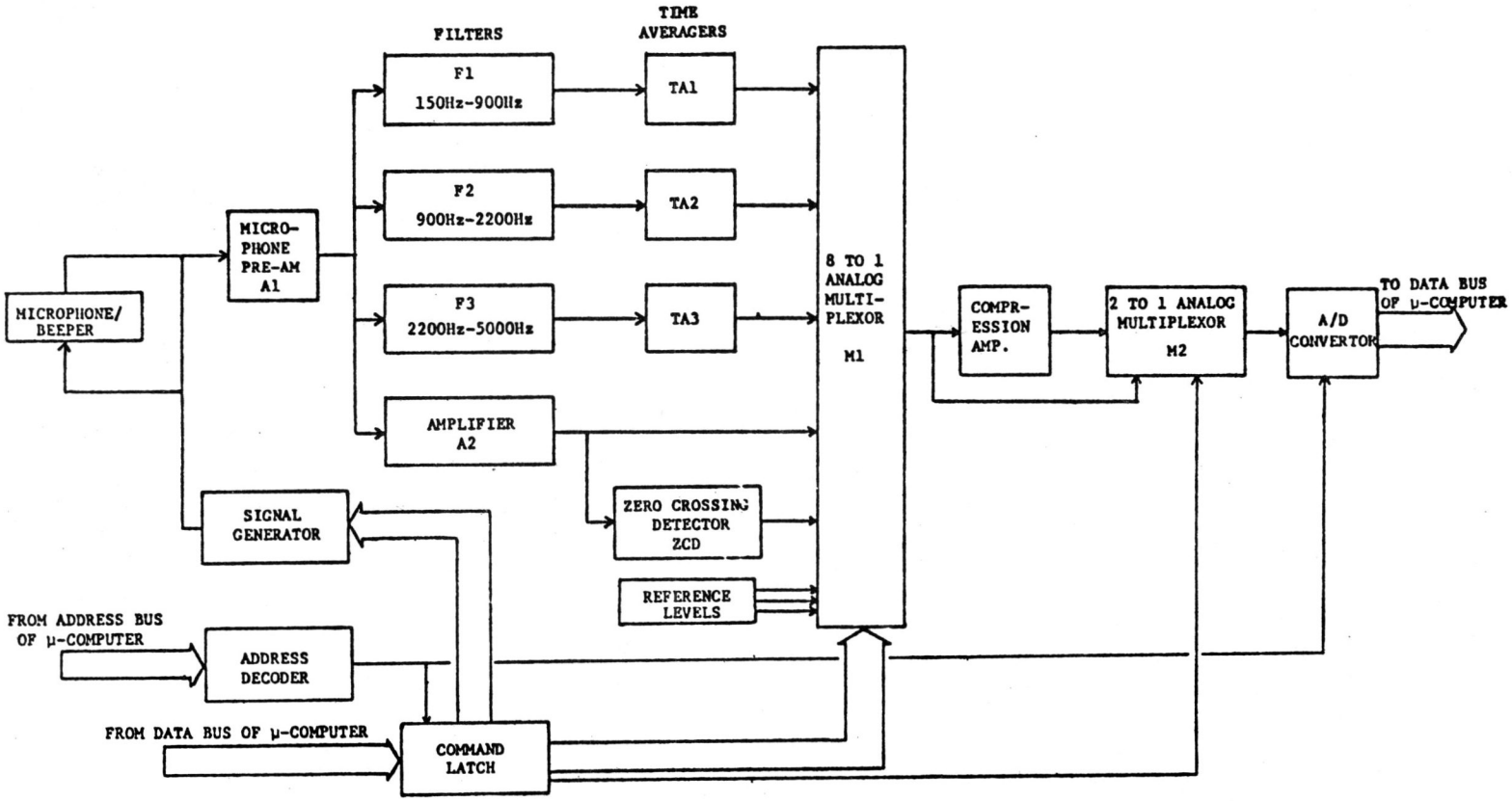


Figure 1: Schematic Diagram

repeated speech.

The analog to digital conversion of the signal arriving from the multiplexer M2 is performed by a 6-bit ramp up-type converter using a ripple counter driven by the system clock and a voltage comparator. The ripple counter generates an increasing voltage to one input of a comparator, the other input being the processed speech signal. The comparator changes state when its inputs are equivalent and stops the counter when its contents are available to the system data bus as a digital representation of the converted analog signal.

SYSTEM SOFTWARE

A) Port Control

Program control of the speech collection and conversion is attained through a word placed on a single input-output port configured as in fig. 2. The three lowest bits of this word select the input at the multiplexer M1 for conversion. Bit 3 of the control word passes to the multiplexer M2 where it is used as a switch control for possible use of the compression amplifier. Bit 4 is used to activate the signal generator to give the beep cues to the speaker and thus delineate the speech window. It may also be used in calibration and testing as the signal generator output contains frequencies in each

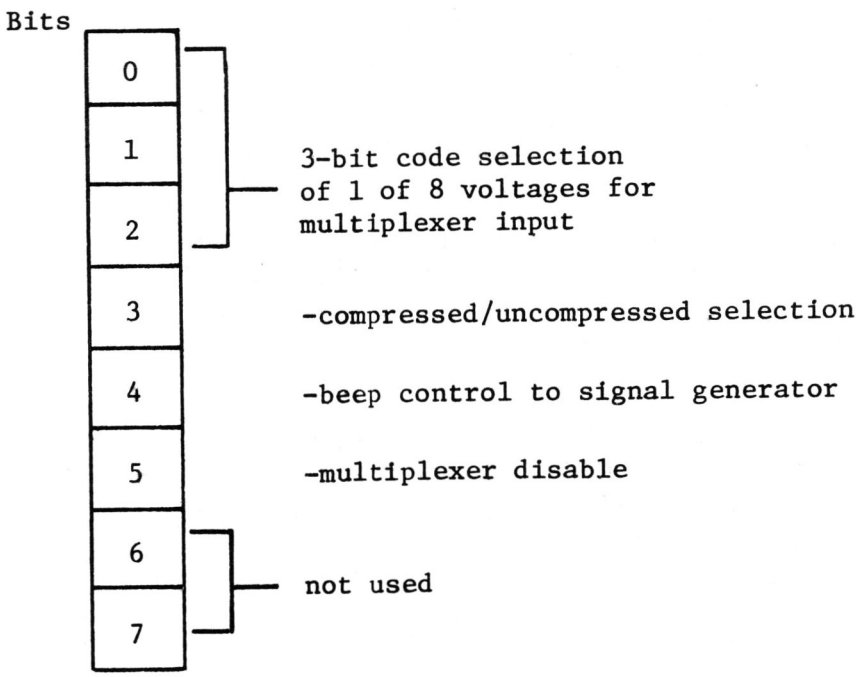


Figure 2: Output port

of the filter passbands. The only remaining output port bit which is used is bit 5 which disables the multiplexer M1 when the lowest three bits are changed. The input port provides the results of the analog to digital conversion in bits 0 to 5 (fig. 3). Bit 6 is not used and is always zero and bit 7 signifies the completion of the analog to digital conversion when the conversion counter is stopped.

| | |
|---|---|
| Bits 0 - 5 | 6-bit analog to digital conversion output |
| Bit 6 | Unused |
| Bit 7 | Converter status |

Figure 3: Input port

B) The Recognition Process

Recognition of spoken words is achieved in two phases. Firstly, a vocabulary is provided as a set of templates for the spoken words together with a table of corresponding written words which are used to identify the result of subsequent recognition. Throughout the duration of the speech window, as indicated by the beeper cues, samples of the spoken word are taken at regular intervals and stored in a buffer. This temporary storage contains consecutive measures from the three filters and the ZCD realized as integers ranging from 0 to 63 resulting from the 6 - bit analog to digital conversion. Typically, with a 1.5 sec speech window, samples are collected every 10 ms which approximates the fundamental period of the male voice. The 600 data samples in the buffer are analyzed in order to extract 64 representative samples which are transferred to another area of memory reserved as the vocabulary table. This is a $n$ x 64 array where $n$ is the size of the vocabulary.

When the vocabulary is in storage the recognition phase can be started with vocal inputs collected and analyzed exactly as before. The 64 data samples selected from the raw speech buffer are moved into a test word buffer where each byte is compared to the corresponding byte of a vocabulary word. A minimum distance classification is performed between the word under test and each of the vocabulary words. The sum of the absolute differences in each byte-pair are stored in an n-element vector which is finally examined to determine its minimum value entry. The position of this value corresponds to the row number of the spoken vocabulary table and thus is used as an index to the written vocabulary table. This is the result of the recognition and it is displayed. Optionally the sum-of-differences vector can also be output from which the word proximities can be ranked.

The analysis of the buffer word starts by locating the word boundaries. When the sum of the processed outputs from the filter banks for each sample exceeds a threshold this is taken as the word beginning provided that the ZCD data is also below another threshold value [5]. If the latter is above its threshold the beginning point is lowered to an earlier sample with a level below its threshold.

A similar technique with movement in the opposite direction finds the end of the word. The use of the zero crossing detector is useful for words with less distinctive extremities involving the unvoiced fricatives. As a stop consonant within a spoken word would give a silence for up to 100 ms, this could be taken as an endpoint and so subsequent samples are examined to ensure there is no rise above threshold in the following 100 ms. If the threshold is exceeded the endpoint is raised to this sample and a further examination is made from this point.

The difference of the endpoints gives the length of the word which is separated into 16 evenly spaced intervals by division to determine 16 representative samples. Sampling by division is done in order to compensate for the variation in the length of a word spoken on different occasions and it is also useful in reducing the required storage and promoting faster recognition. Linear interpolation is used to compute the parameters corresponding to the filter bank and ZCD outputs at each of the samples. The amplitudes of the selected samples are normalized using the mean amplitude to reduce the effect of variation of amplitude of the speech signal.

## SYSTEM PERFORMANCE

This system has been used for research into the parameters of speech recognition and to employ improved techniques to retain an acceptable recognition rate as the vocabulary is expanded. Most of the work done has been with a single speaker and the best recognition rate has been 92% over 500 trials with a 10-word vocabulary. It is noted that our system is designed to adapt itself to recognize words from different speakers accomplished by changing the stored acoustic parameters and that high recognition rates can be maintained or improved. This learning effect can offset an increased error rate when the vocabulary is increased. Further adaptation is easily accomplished by the replacement of the vocabulary. Its possible use in speaker identification has been demonstrated as has its employment in the vocal input of a computer program. It is in this area that most work with this system is proceeding.

## ACKNOWLEDGEMENT

This research was supported by an FCAC grant from the Department of Education, Quebec.

## REFERENCES

1. Rabiner, L.R., "Special Issue on Man-Machine Communications by Voice", *Proc. IEEE*, vol. 64, pp. 403-404, April 1976.

2. Reddy, D.R. and Erman, L.D., "Tutorial on System Organization for Speech Understanding", in *Speech Recognition*, D.R. Reddy, Ed., Academic Press, 1975.

3. De Mori, Renato, "Recent Advances in Automatic Speech Recognition", Fourth International Joint Conference on Pattern Recognition, Kyoto, November 1978.

4. Martin, T.B., "Applications of Limited Vocabulary Recognition Systems", in *Speech Recognition*, D.R. Reddy, Ed., Academic Press, 1975.

5. Shafer, R.W. and Rabiner, L.R., "Parametric Representations of Speech", in *Speech Recognition*, D.R. Reddy, Ed., Academic Press, 1975.