Proceedings of the Technical Sessions, 20 (2004) 1-5
Institute of Physics – Sri Lanka

Using time and frequency domain information for the development of adaptive voice control devices

E.P.N.S. Edirisinghe and D.U.J. Sonnadara
Department of Physics, University of Colombo, Colombo 3

ABSTRACT

A simple voice recognizer was implemented with the aim of building voice control devices by utilizing a cascade LPC with adaptive fuzzy interface system (FIS). The accuracy of the voice recognizer depends on the complexity of the voice commands, the number of frequency components used, and the weights that are utilized in FIS. The quality of the voice recognizer depends both on the time and the frequency resolution of the digitizing hardware. With a limited set of sounds the technique discussed in this paper can be used to produce a speaker dependent voice control system. The performance of the system improves with time due to the adaptive nature of the system. The highly flexible backend (fuzzy interface) enables usage of the device in noisy industrial environments.

1. Introduction

Speech, being the dominant source of communication, has played a crucial role in the evolution of the human race. Due to the rapid advancement in the technology and the ever-increasing complexity of electro-mechanical devices, research is carried out to develop methods that ease the human-device interactions. Today, electronic devices have taken over a majority of the work carried out by humans a decade ago. However, the communication between electronic devices and humans is still at its infancy. The work presented in this paper was carried out to develop a suitable methodology to increase the ease of communication between humans and electronic devices.

During the last twenty years many scientists have proposed various techniques to increase the effectiveness of speech recognition. So far, most of the research related to analyzing sound signals is restricted to one domain, either the time domain or the frequency domain. For example, a technique such as Linear Predictive Coding (LPC) [1] which is generally accepted for speech recognition, completely ignores the frequency domain information. On the other hand frequency domain techniques such as Fourier transforms, completely ignore the time domain information. If information encoded in both time and frequency domains is utilized in interpreting speech signals, it could be possible to achieve better accuracy in the interpretation. This hypothesis is supported by the natural functionality of human auditory systems [2].

A system based on clustered linear predictive coding coupled to a Fuzzy logic interface [3] at the backend was developed to maintain acceptable tradeoff between the time domain resolution and the frequency domain resolution. The objective of this work was

to test the effectiveness of the above technique in building voice command control systems that respond to simple voice commands.

2. Methodology

The study was carried out by concentrating on pre-recorded digitized sound files stored in RIFF WAVE (.WAV) format. The structure of these files was determined and the raw data corresponding to the speech signals were extracted. To reduce the number of data points, a sampling rate of 11 KHz was used in recording these files. Each sample was normalized by converting into a logarithmic scale. This method which is based on the functionality of actual human auditory system has been recommended elsewhere [2].

$$x(i) = 10\log\big(x(i)\big) + 10\log\left(\frac{x(i) + x(i+1)}{x(i)}\right)$$

The first task was to feed the input signal to the LPC in such a way as to reduce its complexity, enabling the LPC to interpret speech signals accurately. This was achieved by clustering signals into a set of frequency bands, based on the *mal* frequency scale. Pre-processed input signals were first passed through the filter banks [4]. The frequency bands (equal size in the mal frequency scale) were selected so that the center frequency of each band was linear in the mal frequency scale. The optimum number of frequency bands was searched for by minimizing the error while maintaining an acceptable processing speed. Ten LPC coefficients were computed for each of the frequency bands and a fuzzy interface system was utilized for decision making at the rear end.

To analyze the LPC data, a Mamdani-type Fuzzy inference system (FIS) was utilized. The FIS consisted of 40 input nodes, 160 input membership functions and 4 output membership functions. Decision-making was achieved with the aid of 16 decision rules. Figure 2 illustrates a simplified block diagram of FIS. It was observed that each of the LPC coefficients follow a Gaussian distribution. Thus, a Gaussian membership function was used as the input membership function. Initially, known signals were fed to the clustered LPC network and output was placed in a database. Based on information gathered from these data, assuming a Gaussian distribution, the distribution parameters of each of the LPC coefficients were calculated. These LPC coefficient distributions were used as input membership functions for the fuzzy interface system.

The developed system is an adaptive one. Once the system identifies a command, it checks the validity of the response by performing a hypothesis test with its stored distribution parameters. If satisfactory accuracy is achieved, the system updates the database with the new data and retains a record of the last data set for each command it responds to. For each command, after receiving 10 new data values, it constructs a new input membership function based on the updated database and verifies the performance of FIS for each command. If the overall system error is reduced, it retains the new FIS for future use.

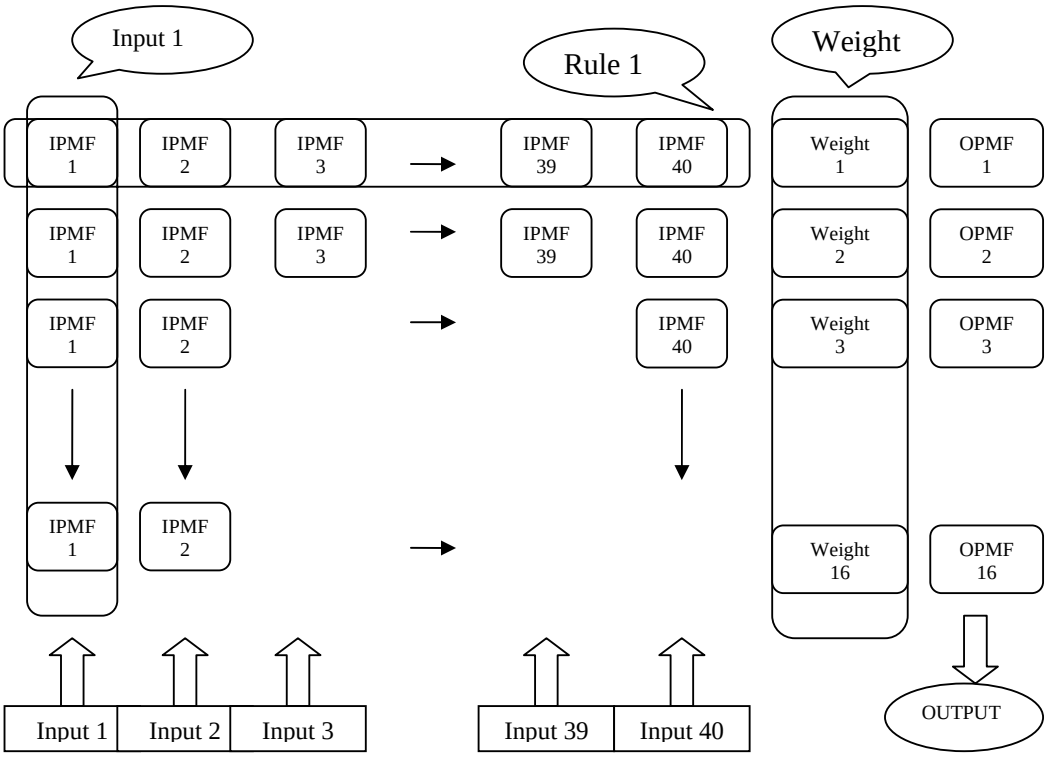

Figure1: Block diagram of FIS

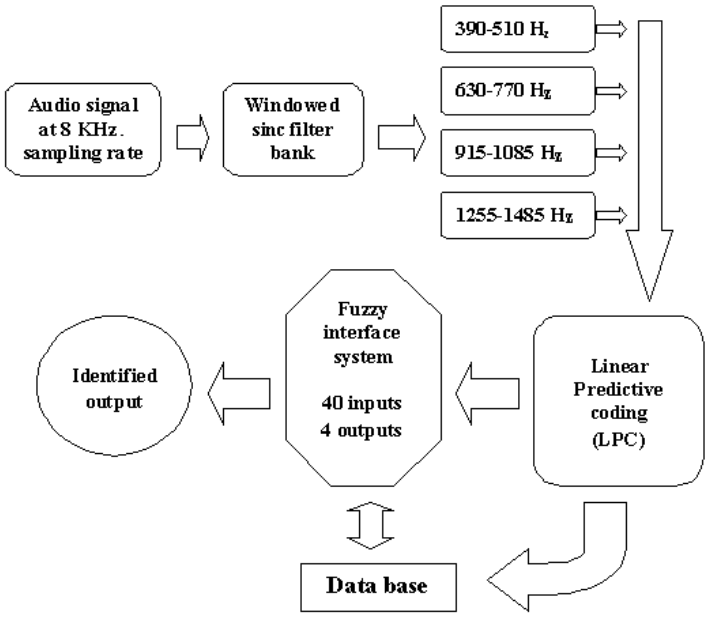

Figure 2: Block diagram of the

3. Results and Discussion

The preliminary results showed that the frequency bands 390-570Hz, 630-770Hz, 915-1085Hz and 1255-1485Hz are adequate to retain all significant information in simple voice commands. Increasing or decreasing the number of bands resulted in adverse effects in the performance. Increasing the number of bands slows down the process while decreasing the number of bands diminishes the accuracy. However, it was noted that the optimal range varies from person to person.

Experimental work revealed that in general, the LPC coefficients follow a Gaussian distribution. Thus, the mean and standard deviation of the sample were computed by observing the sampling distributions. The distribution for the LPC coefficients was computed by assuming a Gaussian distribution and it was used as the input membership function for the fuzzy interface system. Figure 3 shows a few resulting distributions for the LPC coefficients for the frequency band 630-770Hz. It was observed that even for the same LPC coefficients, distributions differ significantly from one coefficient to another. For example, the LPC distributions corresponding to four different selected sounds (ON, OFF, ONE and TWO) differ significantly from one coefficient to another. Thus, to attain acceptable sensitivity, each rule in the fuzzy interface system was weighted according to the relative resolution power of the LPC coefficient distributions for each sound.

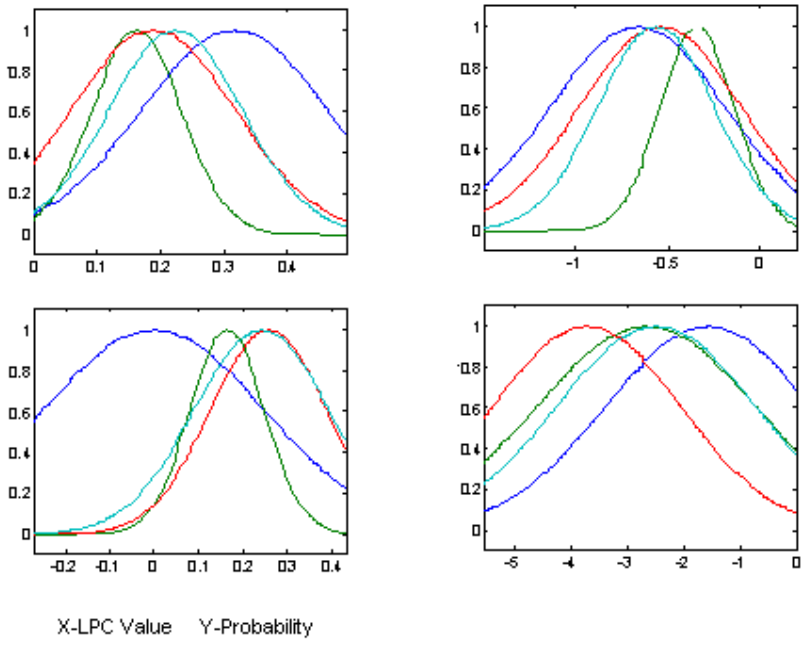

Figure 3: Distributions of LPC coefficients for frequency band 630-770Hz.

A high confidence level was maintained for the database update while the fuzzy level enabled maintaining acceptable decision-making capability. A 95% confidence level

was used as the threshold level and a 1% improvement over existing FIS was used as the criteria for updating FIS.

4. Conclusions

This work shows that the clustered LPC method performs well in recognizing simple voice commands when both time and frequency information is utilized simultaneously. Fourier transform, the technique traditionally used in speech recognition, usually lacks the ability to provide the same resolution in both the frequency and the time domain. Even though windowed Fourier transform provides this up to a certain extent it does not provide the same resolution for different frequencies.

The presence of noise is a major concern for the performance of the system. This makes a critical impact on the adoptability of the methodology in the development of simple voice control devices. Since noises are random in nature it is important to know where the noise originates. Even though clustering the input signal does not enable the localization of noise sources, it enables band limitations in the frequency domain improving signal to noise ratio. Also, the usage of FIS improves the error bandwidth while retaining high decision making capabilities. This also improves the system performance.

The main aim of the work was to search for a technique that enables the development of voice control devices. Thus it was important to construct a system having a unique response with a high level of accuracy. This can be achieved by decoding both time and frequency domain information when interpreting voice commands.

References

1. Press, W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P, (1992) *Numerical recipes in C*, Cambridge University Press
2. Stuart I.F. (1996). *Human Physiology*, 5th edition, McGraw Hill Co. Inc.
3. Zadeh L.A. (1998). *Fuzzy Logic Toolbox User's Guide.* California: The Math Works, Inc.
4. Smith, S.W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing*. California: California Technical Publishing