# DIGITAL SIGNAL PROCESSING ALGORITHMS FOR AUTOMATIC VOICE RECOGNITION



Final Report

NASA/ASEE Summer Faculty Fellowship Program - 1987

Johnson Space Center


| | |
|---|---|
| Prepared by: | Nazeih M. Botros, Ph.D. |
| Academic Rank: | Assistant Professor |
| University & Development: | Southern Illinois University<br>Department of Electrical Engineering<br>Carbondale, Illinois 62901 |
| NASA / JSC | |
| Directorate: | Engineering |
| Division: | Tracking & Communication |
| Branch: | Telemetry & Audio |
| JSC Colleague: | Bill Jordan |
| Date: | August 7, 1987 |
| Control Number: | NGT 44-001-800 |

## ABSTRACT

The main objective of this research is to investigate the current digital signal analysis algorithms that are implemented in automatic voice recognition algorithms. Automatic voice recognition means, in simple terms, the capability of a computer (machine) to recognize and interact with verbal commands. In this research I focus on the digital signal, rather than the linguistic, analysis of speech signal. Several digital signal processing algorithms are available for voice recognition. Some of these algorithms are: Linear Predictive Coding (LPC), Short-time Fourier Analysis, and Cepstrum Analysis. Among these algorithms, the LPC is the most widely used. This algorithm has short execution time and do not require large memory storage. However, it has several limitations due to the assumptions used to develop it. The other two algorithms are frequency-domain algorithms with not many assumptions, but need longer execution time and larger storage, consequently they are not widely implemented or investigated. However, with the recent advances in the digital technology, namely the high density memory chips and the ultra fast digital signal processors, these two frequency-domain algorithms may be investigated in order to implement them in voice recognition. This research is concerned with real-time, microprocessor-based recognition algorithms.

## INTRODUCTION

For more than a decade the United States Government, Foreign countries especially Japan, private corporations, and universities have been engaged in extensive research on human-machine interaction by voice. The benefits of this interaction is especially noteworthy in situations when the individual is engaged in such hands/eyes-busy task, or in low light or darkness, or when tactile contact is impractical/impossible. These benefits make voice control a very effective tool for space-related tasks. Some of the voice control applications that have been studied in NASA-JSC are: VCS Flight experiments, payload bay cameras, EVA heads up display, mission control center display units, and voice command robot. A special benefit of voice control is in zero gravity condition where voice is a very suitable tool in controlling space vehicle equipment.

Automatic speech recognition is carried out mostly by extracting *features from the speech signal and store them in reference templates in the computer. These features carry the signature of the speech signal. These reference templates contain the features of a phoneme, word, or a sentence, depending on the structure of the recognizer. If a

voice interaction with the computer takes place, the computer extracts features from this voice signal and compares it with the reference templates; if a match is found, the computer executes a programmable task such as moving the camera up or down.

A speaker-dependent recognizer is the one that is customized to a particular speaker. The template of this particular speaker is stored in the recognizer memory as the reference template; only this speaker can use that recognizer. A speaker-independent recognizer can be used by any speaker assuming that the speaker's language and dialog are the same of the recognizer. An isolated-word recognizer is the one that can recognize only a single word at a time; A pause should be inserted by the speaker between words. A connected-word recognizer is the one that can recognize a string of spoken words, no pauses needed. A recognizer can be built to recognize the word(s) of the speaker or can recognize the identity of the speaker from his spoken words.

At present time most of the commercially available recognizers are speaker-dependent,isolated word recognition, with limited vocabulary. "Current speech recognition technology is not sufficiently advanced to achieve high performance on continuous spoken input with large vocabularies and or arbitrary talkers." One of the factors that limits the performance of the current recognizers is the

efficiency of the recognition algorithms. "Significant research efforts are required in the design of algorithms and systems for the recognition of continuous speech in complex application domains, for speaker-independent operation, and for robust performance under conditions of degraded input."[1]

Several digital signal processing algorithms are available for speech feature extraction. The efficiency of the current algorithms is limited by: hardware restriction, execution time, and easiness of use. Some of these algorithms are: Linear Predictive Coding (LPC), Short-time Fourier Analysis, and Cepstrum analysis. Among these algorithms, the LPC is the most widely used since it is easy to use, short execution time, and do not require large memory storage. However, this algorithm has several limitations due to the assumptions that the algorithm is based upon. The other two algorithms usually need longer execution time and larger storage and consequently they are not widely implemented or investigated.

[1] Extraction from the National Research Council report, December 12, 1984.

## FINDINGS

In this section we discuss four digital algorithms that been implemented in automatic voice recognition. These algorithms extract a certain number of feature from the speech signal. These features carry the signature of the signal. Recognition is achieved by comparing the feature set of the speaker with the reference feature set. A decision rule is implemented to decide on whether the speaker feature set matches the reference set or not.

1. Filter Bank-Analysis of Speech, [4].

In this algorithm the feature set consists of the speech energy within a certain number of frequency bands. The frequency range of the speech signal is divided into bands. This number varies from 5 to as many as 32, and the spacing between the bands is normally linear until about 1000 Hz, and logarithmic beyond 1000 Hz.

The energy within each frequency band is measured as shown in Figure 1. The speech signal passes through a bank of band pass filters; each filter covers a certain frequency band. The output of the pass band filter passes through nonlinear circuit, such as square law detector or full wave rectifier, and a low pass filter. The output of the nonlinear circuit is proportional to the square of the

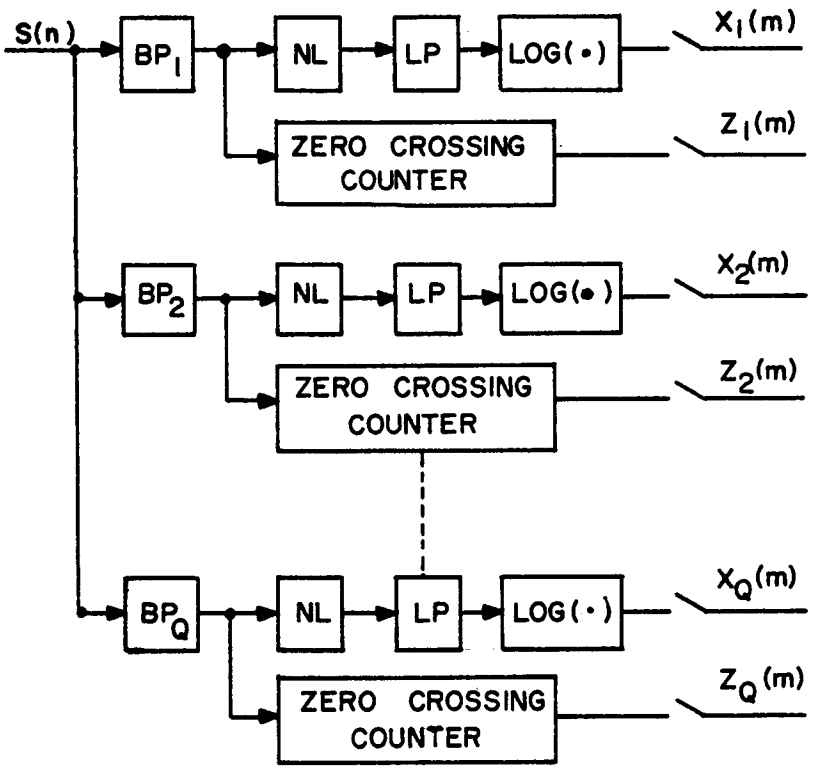


FIGURE 1. FILTER BANK - ALGORITHM.

amplitude of the signal and hence can be taken as a measure of the speech energy in this band. A logarithmic circuit is used to reduce the dynamic range of the intensity signal and the compressed output is digitized with a sample rate of twice the low pass filter cutoff frequency.

The low pass filter cutoff frequency is typically about 20-30 Hz; Accordingly the sample rate of the digitizer is selected to be from 40 to 60 Hz. If the number of the band pass filters is 5 and the sample rate is 40, then the number of features (the energy per frequency band) to represent a 1 second of the speech signal is 200. If we sampled the raw speech signal without using the filter bank, then the number of features will be 1000 for sample rate of 10 KHz. So by using filter bank-analysis the number of features is reduced by a rate of 50 to 1. For many recognizers, this feature set is supplemented by adding the number of times the signal crosses the zero time axis. This number of this zero crossing is related to the frequency pitch of the speech signal.

2. Linear Predictive Analysis, [1-7].

This algorithm is built on the fact that there is a high correlation between adjacent samples of the speech in the time domain. This fact means that an nth sample of speech signal can be predicted from previous samples. The

correlation can be put in a linear relationship and we get what is called Linear Prediction Model. This relationship can be written as:

$$\hat{Y}_n = a_1 Y_{n-1} + a_2 Y_{n-2} + \ldots + a_p Y_{n-p} \qquad \ldots\ldots\ldots\ldots (1)$$

where p is the order of analysis; usually p ranges from 8 to 12. $y_n$ is the predicted value of speech at time n and a's are the linear predictive coefficients. The error $E_n$ that resulted from the above linear relationship is:

$$E_n = Y_n - \hat{Y}_n = Y_n - \left( \sum_{i=1}^{p} a_i Y_{n-i} \right) \qquad \ldots\ldots\ldots\ldots (2)$$

The error $E_n$ is called the prediction error. Setting $-a_i$ as $a_i$, the prediction error becomes:

$$E_n = Y_n + \sum_{i=1}^{p} a_i Y_{n-i} \qquad \ldots\ldots\ldots\ldots (3)$$

$$= \sum_{i=0}^{p} a_i Y_{n-i}, \qquad a_o = 1. \qquad \ldots\ldots\ldots\ldots (4)$$

Squaring the above Equation and taking the average:

$$\overline{E_n^2} = \overline{(Y_n + a_1 Y_{n-1} + a_2 Y_{n-2} + \ldots + a_p Y_{n-p})^2} \qquad \ldots\ldots\ldots\ldots (5)$$

To find the predictive coefficients which give minimum predictive error, the above Equation is partially

differentiated with respect to a's and time average term by term is taken:

$$\begin{bmatrix} r_0 & r_1 & r_2 & \cdots, & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots, & r_{p-2} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots, & r_0 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_p \end{bmatrix}, \quad \text{.....(6)}$$

with

$$r_0 = \overline{Y_n Y_n}, \qquad r_j = \overline{Y_n Y_{n+j}} = \overline{Y_{n-j} Y_n} \quad \text{.....(7)}$$

where $r_j$ is a correlation coefficient of waveform $\{y_n\}$ and $r_{-j} = r_j$ by the assumptions of stationary state of $y_n$ The coefficients $a_i$'s exist only if the matrix in Equation 6 is positive definite. To ensure that this condition is satisfied, $y_n$ is multiplexed by a time window $W_n$. This multiplexing makes $y_n$ exists in a finite interval from 0 to N-1, where N is the interval of the Window; a stable solution for Equation 6 is always obtained. Accordingly, $r_j$ is written as:

$$r_j = \frac{1}{N} \sum_{n=0}^{N-j-1} Y'_n Y'_{n+j} \quad \text{.....(8)}$$

$$= \frac{1}{N} \sum_{n=0}^{N-j-1} Y_n Y_{n+j} W_n W_{n+j} \quad \text{.....(9)}$$

Calculation of the correlation coefficients by window multiplexing is called the correlation method. Some recognizers use the correlation coefficients as the feature set. However for full LPC analysis $a_i$'s are calculated by solving Equation 6.

The LPC model represents an all-pole model. The relation between the input $x_i$ and the output $y_n$ of this system is written as:

$$Y_n + \sum_{i=1}^{p} a_i Y_{n-1} = X_n, \quad \text{.....(10)}$$

Equation 10 is called the auto-regressive process. The system function H(z) can be written as:

$$H(z) = 1/(1+a_1 z^{-1}+.....+a_p z^{-p}) \quad \text{.....(11)}$$

$a_i$'s correspond to the resonance frequencies of the signal and if p, the order of the analysis is selected correctly, these $a_i$'s represent the formants, frequencies at which peaks of the power spectrum of the speech signal occur.

A block diagram representing an algorithm for voice recognition based on LPC analysis is shown in Figure 2. As shown in this Figure, the digitized data is divided into frames each of length N. The distance between consecutive frames is M. If N=M then there is no overlapping between

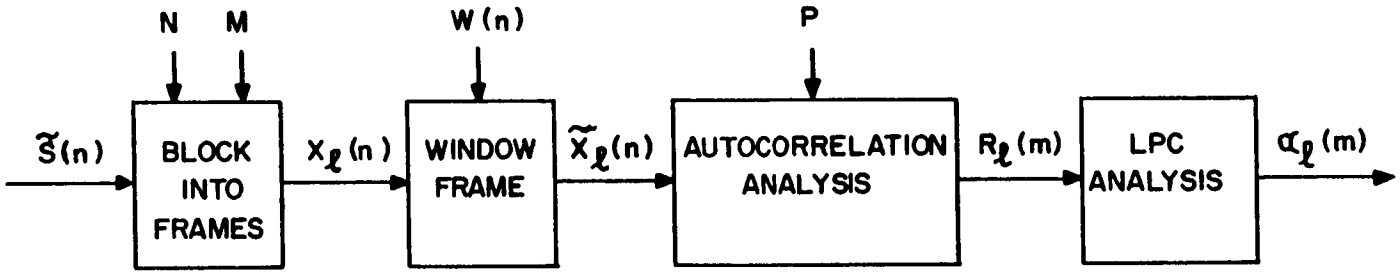


FIGURE 2. LPC-BASED ALGORITHM FOR AUTOMATIC VOICE RECOGNITION

frames; if M<N, then there is overlapping between frames. Typical values of N are from 100 to 500 data points. For large number of N the analysis is called wideband analysis and for small values of N the analysis is called narrow band analysis. Because the rate of speaking of any subject is not fixed, it changes with time, a time wrapping algorithm is used to take this fact into consideration.

Limitations of the LPC analysis

The LPC analysis needs relatively small memory storage and has a short execution time. On the other hand to apply this analysis to any system, the system should satisfy the following conditions:

a. The system is an all-pole system. Speech system does not explicitly satisfies this condition. However, any system can be approximated to all-pole system by increasing the number of poles relative to the zeros in the system.

b. The input to the system is either a single impulse or pure white noise. This is not explicitly true especially in the case of a female voiced sound where the pitch period is generally short.

c. The system is time-invariant. The speech system is time varying system, however using a window can approximate the system to a time-invariant system.

3. Short-Time Fourier Analysis, [2].

In this analysis the domain of the signal is transformed from time domain to frequency domain. Frequency domain analysis is more desirable than the time domain for the following reasons:

a. In the frequency domain the signal is decomposed into its frequency components. Investigation of these frequency components lead to understanding the nature of the signal as well as the effect of the noise on it.

b. The input/output relation of any system in the frequency domain is the product of the Fourier transform of the impulse response of the system and the Fourier transform of the input. In the time domain this relationship is a convolution which is more complicated than multiplication.

c. The autocorrelation function of the system, which is often used to describe the statistical properties of the signal is related in a simple relationship with the power spectrum of the signal.

Since the speech signal is a time-varying signal; the spectrum of the signal changes with time, then Fast Fourier Transform can not be applied directly. Short time Fourier transform should be applied instead of FFT. The short time Fourier transform of a time domain signal, x(m), can be written as:

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} W(n-m)x(m)e^{-jwm} \qquad \text{.....(12)}$$

where W(n-m) is a time window. The window is used here to justify the use of Discrete Fourier transform for a time-varying signal. The Hamming window is the most widely used; it can be written as:

$$W(n) = 0.54 - .46\cos(2\pi n/(N-1)); \quad 0 \leq n \leq N-1 \qquad \text{.....(13)}$$
$$= 0 \quad ; \quad \text{elsewhere}$$

As is shown in Equation 12, short time Fourier includes both convolution process and DFT process. Calculation of short time Fourier by Equation 12 takes a very long execution time. It can be shown that Equation 12 can be written as:

$$X_n(e^{jw}) = e^{-j\frac{2\pi}{N}kn} \sum_{q=0}^{N-1} \{ \sum_{r=-\infty}^{\infty} x_n(Nr+q)\} e^{-j\frac{2\pi}{N}kq} \qquad \text{.....(14)}$$

$$X_n(e^{jw}) = e^{-j\frac{2\pi}{N}kn} \sum_{q=0}^{N-1} U_n(q)e^{-j\frac{2\pi}{N}kq} \qquad \text{.....(15)}$$

where

$$U_n(q) = \sum_{r=-\infty}^{\infty} X_n(Nr+q) \qquad \text{.....(16)}$$

As is shown in Equation 15, the short time Fourier Transform can be written as an FFT process multiplied by : $e^{-j\frac{2\pi}{N}kn}$ ;the execution time is relatively short and consequently Equation 15 is used to calculate the short time

Fourier Transform. After calculation of the short time Fourier Transform, the spectrum can be investigated to extract features, such as formants, that can represent the speech signal.

4. Cepstrum Analysis, [1],[2].

The speech waveform can be approximated by:

$$s(t) = \sum_{n=-\infty}^{\infty} s_0(t-nT) = s_0(t) * [\sum_{n=-\infty}^{\infty} \delta(t-nT)] \quad \ldots\ldots\ldots(17)$$

where $s_0(t)$ is impulse response of the speech generating system and $\sum_{n=-\infty}^{\infty} \delta(t-nT)$ is the pulse train with period T. Applying Fourier Transform to both sides of the above Equation, then:

$$S(w) = S_0(w)\{\sin[(2N+1)\tfrac{1}{2}wT]/\sin\tfrac{1}{2}wT\}^2 \quad \ldots\ldots\ldots(18)$$

Where S(w) and $S_0(w)$ are the power spectra of s(t) and $s_0(t)$ respectively. Taking the logarithm of both sides of the above Equation, then:

$$\log_e S(w) = \log_e S_0(w) + 2\log_e\{(2N+1)\tfrac{1}{2}wT]/\sin\tfrac{1}{2}wT\} \quad \ldots\ldots\ldots(19)$$

The first term on the right-hand side of the above Equation represents a relatively slow change in frequency (the speech generating system) and the second term represents

a relatively high change in frequency with fundamental frequency $2\pi/T$. This means that the above Equation consists of two separable terms with respect to frequency. By taking the Inverse Fourier Transform, we can have two terms; the first term corresponds to the spectrum envelope and the second term corresponds to the pitch excitation. The result of this inverse Fourier transformation is called cepstrum and the variable corresponding to frequency is called quefrency. Figure 3 demonstrates the cepstrum analysis.

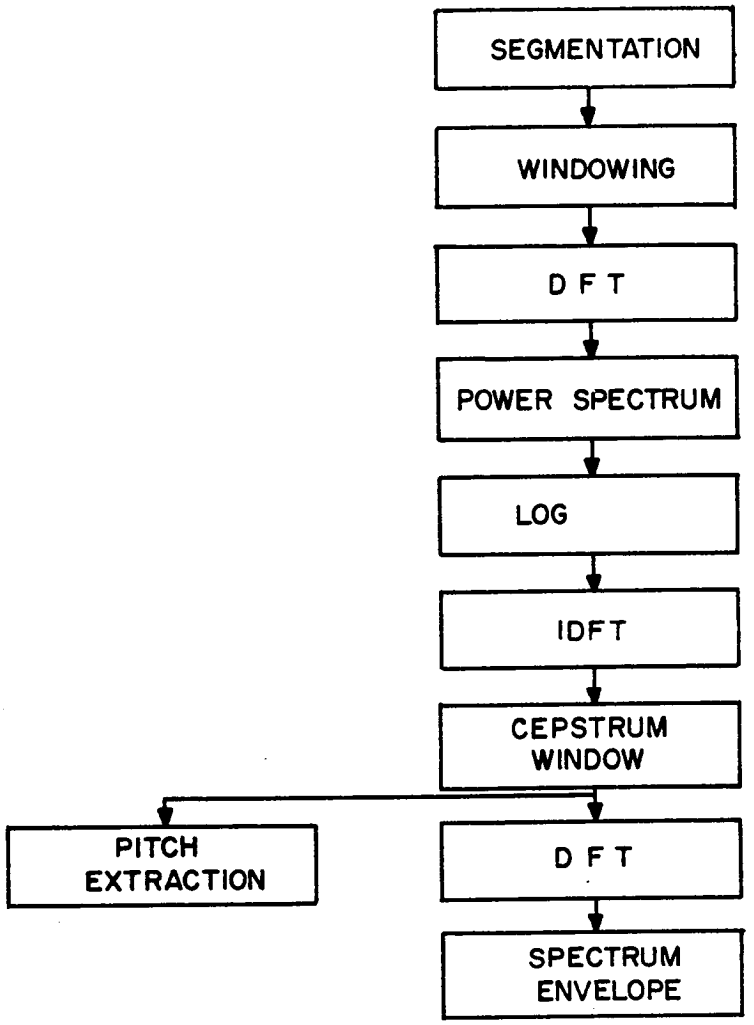


FIGURE 3. CEPSTRUM ANALYSIS OF SPEECH

## CONCLUSION

Four digital signal processing algorithms for automatic speech recognition have been discussed in this research. The Linear Predictive Analysis is the most widely used for its low memory storage requirement and its short execution time. However this algorithm has several limitations. The Short Time Fourier Analysis and the Cepstrum Analysis are frequency-domain algorithms; these two algorithms require large memory storage and their execution time is relatively long. These two frequency domain algorithms do not have as many limitations as of the LPC analysis. The best approach to introduce a high performance algorithm for automatic voice recognition is a combination of LPC analysis and the frequency domain Analysis.

## REFERENCES


1. Saito, S. and Nakato, K., 1985, Fundamental of Speech Signal Processing, Shuzo Saito and Kazuo Nakato, Academic Press, Inc.
2. Fallside,F. and Woods, W., 1985, Computer Speech Processing, Prentice Hall International (U.K.) Ltd.
3. Itakura, F., 1975, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech, Signal, Processing, Vol. ASSP-23, pp. 67-72.
4. Rabiner, L. R.,1978, "On Creating Reference Template for Speaker Independent Recognition of Isolated Word", IEEE Trans. Acoust., Speech, Signal, Processing, Vol. ASSP-26, pp. 34- 42.
5. Rabiner, L. R. and Levinson, S. E., 1981, "Isolated and Connected Word Recognition-Theory and Selected Application", IEEE Trans. on Communication, Vol. comm. 29, No. 5.
6. Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G., 1979, "Speaker-Independent Recognition of Isolated Word Using Clustering Technique", IEEE Trans. Acoust., Speech, Signal, Processing, Vol. ASSP-27, pp. 336-349.