ISOLATED SPEECH RECOGNITION SYSTEM

Mustafa AKTEKİN[1], Ayşen DALOĞLU[2]
[1]TELETAŞ, Araştırma Geliştirme Bölümü Ümraniye/İstanbul, Türkiye
[2]YILDIZ Üniversitesi, Kocaeli Müh. Fakültesi İzmit /Kocaeli, Türkiye

ABSTRACT

In this work, an isolated speech recognition system designed for a digital exchange is introduced. The hardware structure of the system is given and the functions of the each module are explained. Software steps and the algorithms are given, the training phase of the system and to build up the library in the system memory are discussed.

1. INTRODUCTION

Nowadays, digital exchanges are getting to be more powerful and more functional to serve the user. Basic function of a switch is setting up a proper speech communication channel. If we want to implement an operator function as an automated service in the system, we require a speech recognition process anyway.

Telephone services covers some manually operated services by an operator such as alternate billed calls. Alternate Billed Calls (ABC) refer to those calls billed to a number other than that of the originating telephone. This includes to collect calling information from the caller, to make a conversation with called party about the call acceptance, and to proceed the call control under called party wish. In essence, Alternate Calls service has been implemented by TELETAŞ as an Automated Alternate Billing System (AABS). In essence, AABS simulates the customer interaction that would normally be performed by an operator.

AABS has the following main functional steps;

1-Calling party dials a special service calls to initiate AABS service, a special announce, instruct to the calling party to dial the called party phone number.

2-AABS service, establish direct connection between "AABS and called party.

3-After off hook of the called party phone AABS generate a synthetic speech to explain the calling party phone number and service usage procedure.

4-Speech recognition part of the AABS recognize the called party answer and activate a proper AABS function for remaining actions.

5-If, the called party accepts the call, a direct connections established between the calling and called party, but charging information is transferred to the called site.

If, the called party refuse the call, calling party receive an announce about the refusal.

This paper concerned about the speech recognition system employed in the TELETAŞ's Automated Alternate Billing System. The system, basically depends on the isolated speech recognition process.

2. Implemented Speech Recognition System

Any speech recognition system can be implemented on the following functional steps;

-Speech segmentation and parameterisation
-Speech segment identification and recognition
-Word identification

Normally, word recognition require speech segment recognition, and word recognition processes on the sequence. Although, any recognition process requires two logical processes.

a- Training Process
b- Recognition Process

Training process may help us to set up two reference library for the recognition system.

Turkish language speech segments library
Turkish language word definition library

The speech segment library may contain all of the basic Turkish Language speech segments with our parameter set on the vector quantized form. Any extracted segment parameters can be identified by using existing segment library.

Word definition library contains relevant speech segment set to define each word .

3. THE HARDWARE STRUCTURE OF THE SYSTEM

In the design of this speech recognition system, two microprocessors are used. One of them is a digital signal processor as TMS320C15, that realizes the feature analysis process. The other microprocessor, Intel 80C186 is the master controller of this system. 80C186 takes speech samples from PCM channel via a serial port (8920) and writes them to RAM. The data stored in RAM is converted to packets, each containing 80 speech samples. 80C186 writes the packets to a dual port RAM and the digital signal processor reads them. Then, signal processor makes parameter extraction process from the samples. All speech data are sampled at 8 KHz, PCM frequency and analysed for 10 ms window length . Digital signal processor, makes windowing and computes the auto correlation coefficients, the linear predictive coding coefficients (LPC), zero crossing counts. Then, determine the formant frequencies from the LPC coefficients.

Digital signal processor writes these analysis parameters to the dual port RAM. The 80C186 reads the packet of parameters from the dual port RAM, quantizes them by using vector quantization algorithm. The evaluation of these parameters for segmentation and making decisions are realized on the 80C186 part of the system. The hardware structure of the system is illustrated in Fig. 1

4. THE ANALYSIS PROCESS OF THE SPEECH

The software which is implemented on the digital signal processor, makes the feature analysis process. The feature analysis, distills the information necessary for speech recognition from the raw speech waveform. Just as important, it discards information such as; background noise, channel distortion, speaker characteristics and manner of speaking.

Commonly used feature set for recognition is the LPC (Linear Predictive Coding) based feature set. The basic idea behind linear predictive coding is that a given speech sample can be approximated as a linear combination of the past speech samples. By minimizing the sum of the squared differences (over a finite interval) between the actual samples and the linearly predicted ones, a unique set of predictor coefficients can be determined. Linear predictive coding can be readily shown to be closely related to the basic

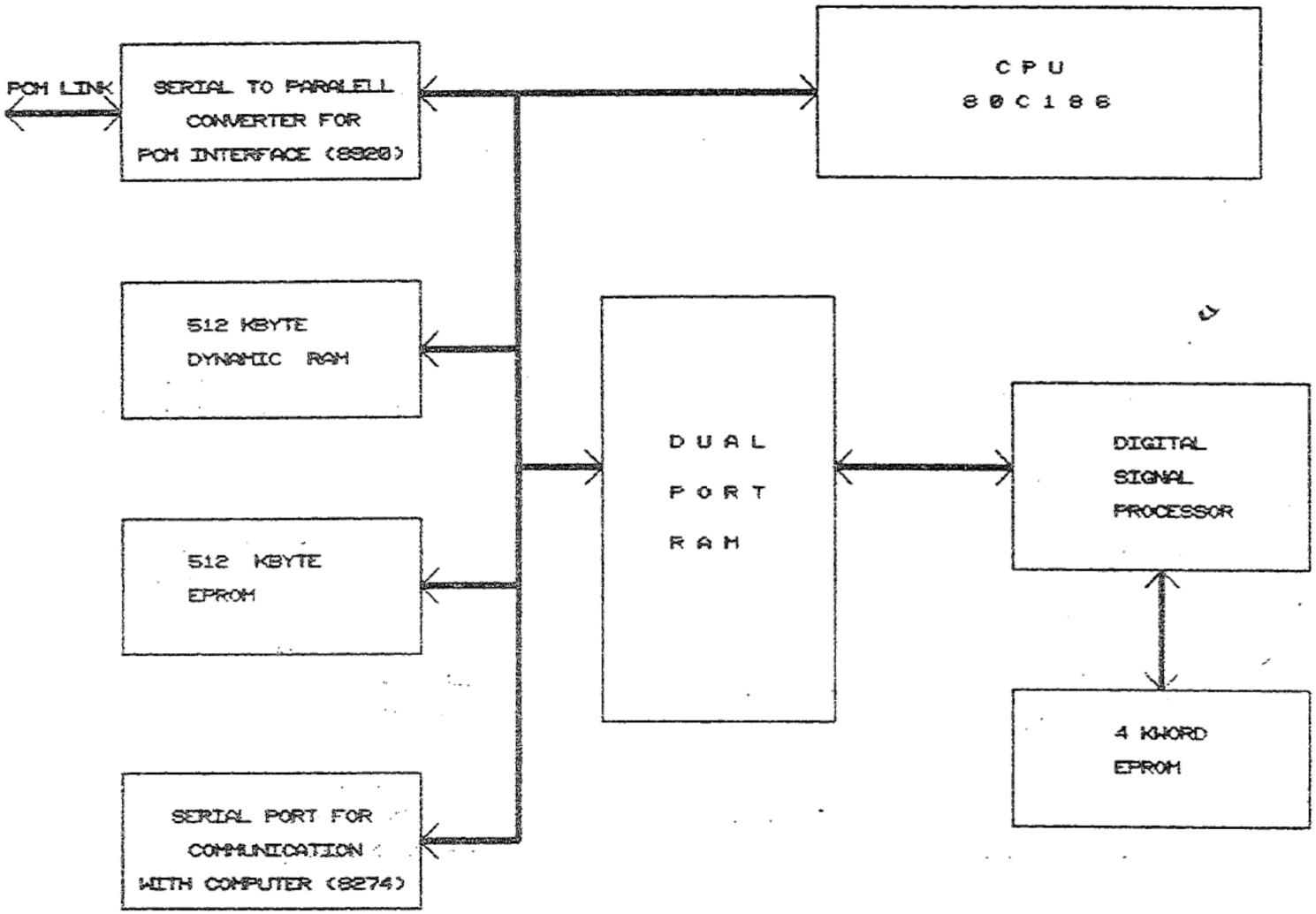

FIGURE 1. THE HARDWARE BLOCK DIAGRAM OF THE SYSTEM

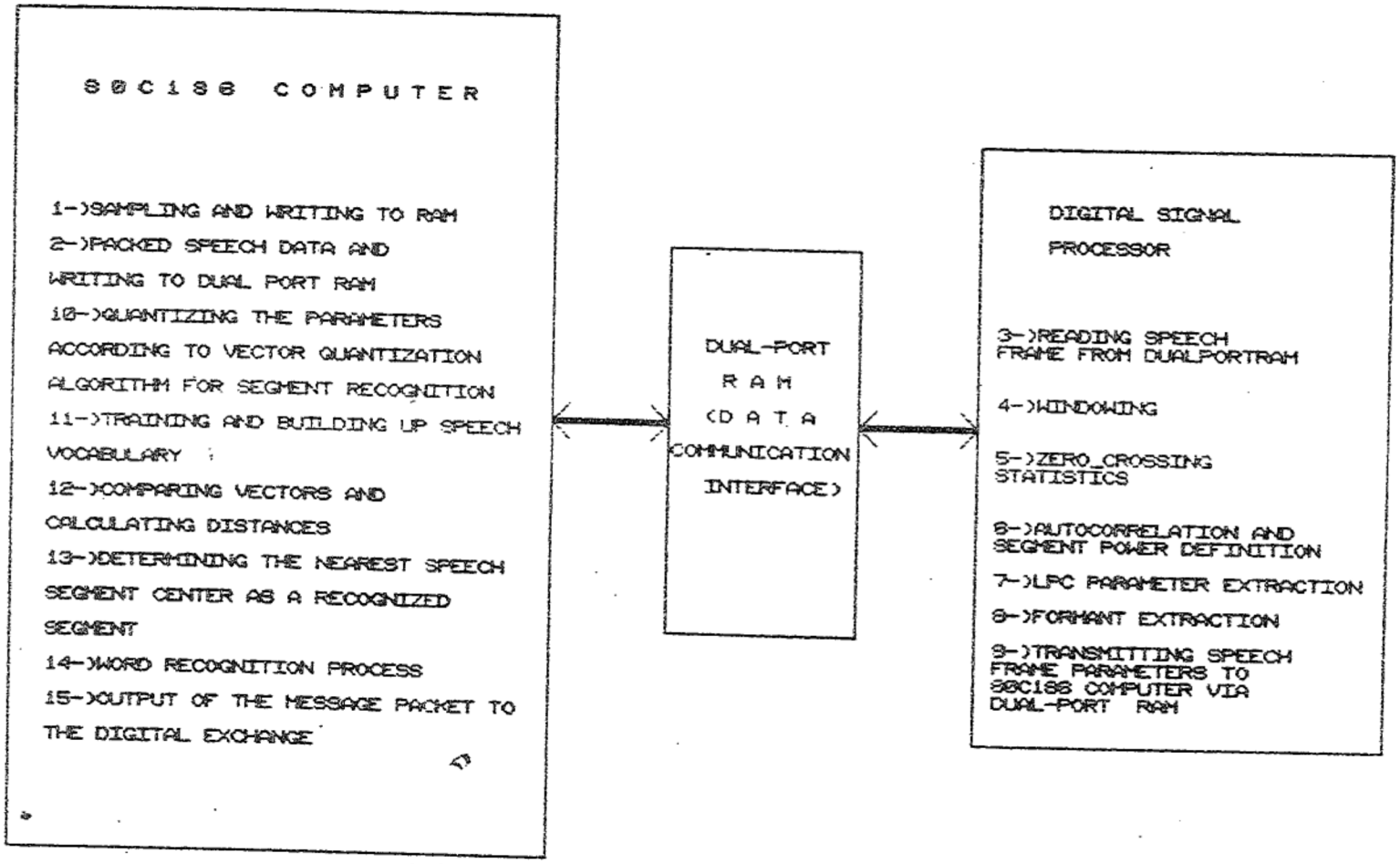

FIGURE 2. THE FUNCTIONAL BLOCK DIAGRAM OF THE SYSTEM

model of human speech production in which the speech signal S_l modelled as the output of a linear, time varying system excited by either quasiperiodic pulses (for voiced sounds) or random noise (for unvoiced sounds). The linear predictive coding method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear, time - varying system.

LPC based feature analysis system is a block processing model in which a frame of N samples of speech is processed, and F feature vector is measured.

To obtain F vector, the speech signal is first pre-emphasized using a fixed first order digital system with transfer function;

$$H(z) = 1 - a.z^{-1}\ ,\ a = 0.95 \qquad (1)$$

giving the signal,

$$\tilde{s}(n) = s(n) - a.s(n-1) \qquad (2)$$

The signal is next blocked into I sample sections (frames) for feature measurement. In order to minimize the effects of the short time duration the analyzing of the speech waveform, a smoothing window, w(n) is applied to the data packet to taper the speech samples to zero at the end of the frame, giving the windowed signal,

$$\tilde{X}_l(n) = X_l(n)\ .\ w(n) \qquad (3)$$

A typical smoothing window, used in LPC analysis systems is the Hamming window defined as,

$$w(n) = 0.54 - 0.46\ \mathrm{Cos}\ \left(\frac{2\pi n}{N-1}\right) \qquad (4)$$

The next step in the feature analysis software is to perform an autocorrelation analysis of the windowed frame of data, giving.

$$R_l(m) = \sum_{n=0}^{N-1-|m|} \tilde{X}_l(n).\ \tilde{X}_l(n+m)\ ;\ m = 0,1,..p \qquad (5)$$

Where, p is the order of the analysis system. The set of the equations can be expressed in a matrix form as,

$$\begin{bmatrix} R_l(0) & R_l(1) & R_l(2) & R_l(p-1) \\ R_l(1) & R_l(0) & R_l(1) & R_l(p-2) \\ R_l(2) & R_l(1) & R_l(0) & R_l(p-3) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ R_l(p-1) & R_l(p-2) & R_l(p-3) & R_l(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} R_l(1) \\ R_l(2) \\ R_l(3) \\ . \\ . \\ . \\ R_l(p) \end{bmatrix} \qquad (6)$$

By taking advantage of the Toeplitz nature of the coefficients matrix (6), several efficient recursive procedures have been devised for solving this system of equations. In the design of our feature
extraction software, to obtain LPC parameters ($a_j = a^{(p)}_j$, $1 \le j \le p$), Durbin's recursive solution is used.

Linear predictive analysis of speech has several advantages when applied to the problem of estimating the formants for voiced sections of speech. Formants can be estimated from the linear prediction parameters. Given the LPC coefficients a_j, $j = 1,2, ..., p$ computing $A_k = DFT \{ 1, a_1 , a_2, ..., a_p , 0 .....0\}$ to obtain the discrete Fourier transform of the inverse filter, simple minimal picking on $|Ak|^2$ for each frame gives the raw data from which formant frequencies can be estimated.

Finally, in the parameter extraction process, the short time energy of the speech signal and zero crossing rate of the speech data are calculated by the digital signal processor to be used in decision making stage.

5 THE EVALUATION OF ANALYSIS PARAMETERS

Analysis parameters are coded by using vector quantization algorithm for the segment recognition process. In vector quantization, we need to determine the reconstruction levels r_i and corresponding cells C_i. A list of reconstruction levels is called a reconstruction codebook or a codebook. If there are L reconstruction levels in the list, the list is said to be an L - Level codebook. A codebook is normally generated by a training procedure which minimizes the average distortion resulting from coding a suitably long sequence of vectors. We suppose that we have M training vectors denoted by f_i for $1 \le i \le M$. Since, we estimate L reconstruction levels from M training vectors, in that calculation we assumed $M >> L$. The reconstruction levels r_i are determined by minimizing the average distortion, which is defined by,

$$D = \frac{1}{M} \sum_{i=1}^{M} d(f_i, \hat{f}_i) \qquad (7)$$

The algorithm steps can be summarized such as,

1) We begin with an initial estimate of r_i for $1 \le i \le L$.

2) We then classify the M training vectors into L different groups or clusters, corresponding to each reconstruction level using the equation;

$$VQ(f) = r_i \text{ , if and only if } d(f, r_i) \le d(f, r_j),\ j \ne i,\ 1 < j < L \qquad (8)$$

This can be done by comparing a training vector with each of the reconstruction levels and choosing the level that results in the smallest distortion.

3) A new reconstruction level is determined from the vectors in each cluster. Let us suppose, f_1 for $1 \le i \le M_1$ are M_1 training vectors quantized to the first reconstruction level r_1. The new estimate of r_1 is obtained by minimizing ;

$$\sum_{i=1}^{M1} d(f_i, r_1) / M_1 \qquad (9)$$

4) A new estimate of all other reconstruction levels r_i for $2 \leq i \leq L$ is similarly obtained.

5) This iterative procedure can be stopped when the average distortion D does not change significantly between two consecutive iterations.

The basic idea, that by using vector quantization; it is possible to set up a separate codebook for each analyzed speech segment. During the training phase, each distinct speech segments are determined and quantized by a vector, called 'Speech Segment Center'.in the vocabulary. Segment recognition is then simply a question of which speech segment center in the segment library fits the analyzed segment vector. The vectors that are generated in a parameter extraction step, are compared with speech segment centers in the vocabulary, then their distance measures are calculated. In the making decision stage, the nearest speech segment center in the library is determined as a recognition segment. Then, extracted segment information is stored for word recognition process.

The following each other similar speech segments are given with the unique common identity. If, the segment identities exceed limits of the speech segment center, the forward alternation in voice signal is observed and directed new speech segment center is determined. The origin point of this alternation and the destination point are registered. The silence segments are deleted. But, spoken word is determined by using stored recognized speech segment centers in two consecutive silence segment as a distinct word.

In the word library, the conception which is correspond to speech segments is evaluated as the recognized word. The word causes a specific message flow, at the digital exchange to initiate a proper process.

6 TRAINING and BUILDING UP THE LIBRARY

Training system has similar hardware architecture with the recognition system. Training system has two phases ;

a-Speech segment library generation phase

b-Word library generation phase

6.1 Speech Segment Library Generation

Speech segment library requires manual training on the following sequence;

1) All of the vowel sounds of the speech are analyzed end classified by manually. The vocal tract maintains a relatively stable configuration during the production of Turkish vowel sounds. Turkish vowels are characterized by a negligible nasal coupling, and by radiating from the mouth. This feature gives us an opportunity to introduce vowel speech segment directly from the speech channel to the system. Each vowel is classified with a set of parameter called 'segment parameter set'. All of the test subject produce their own characteristic 'segment parameter set'. Gravity center of the total training results for each voiced sound is called 'segment parameter center'.

2)Consonant sounds are classified as 'fricative consonants', 'stop consonants', nasal consonants', 'glides and semivowels'. Consonant sounds speech centers are classified by using voiced sounds as being before or after the vowels. Training program use existing vowel speech segment information to determine the transition. This speech transition gives a similar transition over from one voiced speech segment center trough the consonant sound segment center, or vice versa. Each phoneme may cause to be produced a set of phonetical speech segment center by the system . Training system may capture the phonetic speech segment centers and the sound travailing route on the phoneme via two or more speech centers.

The speech segment parameter centers are analyzed and classified by manually for each phonemes. From the vowel sounds trough the longest phoneme in turkish language may

create the personnel speech segment library. After using a number of test persons for training program, the system determine the boundaries for each speech segment center. This information is sufficient enough to build the 'speech segment library'.

6.2 Word Library Generation and Usage

Word library is a set of information about each word which is generated many test person's articulation. Training program stores speech segment center addresses sequentially for each word. Various persons can be used to determine the pronunciation difference for each word. This teaching process gives to the system general traveling trace via the speech centers for each word. TELETAŞ's system aims to recognize a distinct word, which means to take an action this way or the other way. Remaining process will be done on the high level system. After recognition of the word a special message packet will be prepared to activate the proper process .

7 CONCLUSION

This work aimed a practical purposes on our application. This process requires a limited number of recognition in the practice. But implementation has been done for a larger word dictionary. In the future work, the recognition system feature can be extended toward the sentence recognition.

8 REFERENCES

1. H. Abut, R.M.Gray, G. Rebolledo, "Vector Quantization of speech and Speech - Like Waveforms " IEEE Transactions on acoustics, speech, and signal processing, vol, ASSP - 30, June 82

2. Y.Linde , A.Buzo, and R.M Gray, "An Algorithm for Vector Quantizer Design ", IEEE trans. commun., vol. com-28, jan.1980, pp 84-95

3. L.R Rabiner and R.W Schafer , "Digital Processing of Speech Signals. Englewood " Cliffs, n.j Prentice _ Hall ,1978

4. J.D. Markel , A.H. Gray , H. Wakita , "Linear Prediction of Speech Theory and Practice" , Speech Communications Research Laboratory, Inc. september,1973.

5. L.R. Rabiner , S.E. Levinson , "Isolated and Connected Word Recognition."IEEE Transactions on Communication, Vol. Com.-29,5,
May 1981