The 9-th convention of Electrical & Electronic Engineers in ISRAEL, Apr. 1975
IEEE Israel 1975

A/2/4

A MINICOMPUTER IMPLEMENTATION OF AN ISOLATED-WORD RECOGNITION SYSTEM

Isaac Engel
Armament Development Authority
Israel Ministry of Defense

David Malah
Faculty of Electrical Engineering
Technion-Israel Institute of Technology

Abstract

The paper describes the implementation of a computationally efficient isolated-word recognition system (IWRS) on a Nova 2 minicomputer. Since the whole recognition process is done by software, great emphasis has been put on reducing the computation load, without degrading the recognition performance.

Time scale compression (TSC) of all input utterances, to the same duration, is performed at the preprocessing phase. This is done by means of a recently developed efficient algorithm, which requires (in this application) only one multiplication and two additions per input sample. By using compression factors of up to 3 a comparable reduction in computation is achieved in the later phases of the recognition process.

In the parameter extraction phase 6 partial correlation coefficients (PARCOR) are extracted per 15 msec segment of the compressed input signal, so that only 102 parameters are necessary to represent each input word of up to 1 second duration.

The fact that all parameter vectors, representing the input utterances, have the same dimensionality avoids the need for using computation consuming methods for time normalization, such as Dynamic Programing (DP), which usually combines time normalization and classification. In this system classification is done by means of the simple Chebyshev distance function.

In 360 recognition tests, with the first ten Hebrew digits spoken by 3 speakers, the system achieved a recognition score of 99.1%. Additional 240 tests have been performed under noisy conditions (S/N=20dB) with a score of 97.9%.

For better evaluation of the proposed system, the same tests have been performed on a reference system which does not use TSC at the preprocessing phase, and applies DP for classification. The reference system required 3 time more computations but achieved the same score in the noisy case and only 98.3% in the quiet-environment case.

I. Introduction

In typical isolated-word recognition systems (IWRS), the recognition process of an unknown input utterance is divided into three main processing phases. These are[1] : preprocessing, parameter extraction and classification.

The preprocessing phase includes operations such as amplitude normalization, untterance boundaries determination, removal of silent or redundant segments, and preemphasis of the input signal. In the parameter extraction phase a parametric representation of the input signal is obtained. These two phases can be performed by a computer system or by a special purpose hardware, but the next phase classification is always implemented on a computer system.

Classification is performed by time normalization and comparison of the parameter vector, which represents the unknown input utterance, with stored reference vectors which represent the reference utterances of the system vocabulary. The unknown utterance is classified as that utterance in the reference library to which it is most similar or nearest, in terms of an appropriate similarity or distance measure.

The paper describes the implementation of a computionally efficient isolated-word recognition system (IWRS) on a Nova 2 minicomputer. Since the whole recognition process is done by software, a great emphasis has been put on reducing the computation load, without degrading the recognition performance.

In this work an approach is proposed for significantly reducing the recognition time. The approach is to perform time scale compression (TSC) of all input utterance, to the same duration, at the preprocessing phase. This is done by means of a recently developed efficient algorithm[2]. By using time scale compression factors of up to 3 a comparable reduction in computations is achieved in the later phases of the recognition process.

The TSC and other measures for reducing the amount of computations at each phase of the recognition process are detailed in the sections II-IV.
The flow chart of this system is given in Fig. 1. (in accordance with the general scheme in [1]).

For evaluation of the proposed system, a comparative system without TSC was developed. This system, which is described in section V, applies dynamic programming for calssification.
Recognition tests of the same utterances were performed with the two systems. These experiments and the results are given in section VI.

The comparison between the performance of the two systems and conclusions are contained in the last section.

II. Preprocessing Phase

A. Utterance Boundaries Determination.

The operations which are described in this section are the first step in the preprocessing phase.

The software written for the system enables two modes of operation. In the "real-time" mode the output signal from the microphone is amplified, band limited to the range 200-3200 Hz, sampled at a 10 KHz rate using a 12 bit A/D converter, and stored in digital form in the computer's memory for further processing. In the other mode the utterances are first recorded by an analog cassette recorder and the digital data base is prepared as detailed above.

only input samples which belong to the same quasi-stationary interval are being weighted in. That is because from (2) the range of the imput samples which are being weighted is $(m-1)N_p=200$, or 20msec for sampling rate of 10KHz, and the intervals of signal quasi-stationarity are typically in the range of 20-40 msec.

The value of the compression factor C varies from one input utterance to the other. It is recomputed for each input utterance in order that all compressed utterances be of equal duration.

The compression factor C_j for the j-th utterance is given by [3].

$$C_j = [L_j - (m-1)N_p] / L_c \qquad (6)$$

where

$$L_c = L_{co}[1+2/(3N_p)], \qquad (7)$$

L_j is the number of samples in the j-th input utterance to be compressed, and L_{co} is the desired number of samples for the compressed utterance. Using L_c instead of L_{co}, in (6), is based on the result obtained in [2] that the relative error in Cj is limited by $0.5(C_j-1)^2/(C_jN_p)$. Limiting C_j to 3 yields, in the worst case, that L_c of (7) should be used.
In the implemented system $L_{co}=3000$, Lco was determind such that it contains an integral number of analysis segments to be used for parameter extraction in the next phase of recognition process, and so that the maximum compression factor to be used is limited to 3 (approximately). As utterance duration L_j, were in the range of 0.5 to 0.9 sec, the compression factors used were in the range $1.66 \leq C \leq 3$. C_j should be limited to 3, because this value was found to be the perceptual limit of compressed speech [5].

For m=2, a particularly simple choice for h(t) which satisfies the required constraints [3], is the shifted symmetrical triangular function

$$\hat{h}(t) = t/(N_pT) \qquad 0 \leq t \leq N_pT$$
$$= 1-\hat{h}(t-N_pT), \quad N_pT \leq t \leq 2N_pT \qquad (8)$$

Hence the general expression (2) for computing an output sample Y, is replaced in this case by

$$Y = h_o(k)[s(k+N_p) - s(k)] + s(k) \qquad (9)$$

where

$$h_o(k) = k/Nsc, \quad k=1,2,\ldots Nsc \qquad (10)$$

replaces $\hat{h}(t)$.

With this modification, only one multiplication and two additions per output sample are required. The total amount of computations required for applying the TSC algorithm constitutes, therefore, of $L_{co}=3000$ multiplications and $2L_{co}=6000$ additions, for each input utterance.

III. Parameter Extraction Phase.

There are three main requirements for parametric representation for utterance recognition. The first, that it contains only relevant information about the utterance, without redundancy. The second, that this information enables correct discrimination of the various utterances. The third, which is particularly important for the implemented system, that the amount of computation is minimal.

The determination of the parametric representation for the utterances in the implemented system was based on the works of Ichikawa [6] and White and Neely [7].
Ichikawa [6] tested the folowing parametric representations: smoothed logarithmic power spectrum, cepstrum, autocorrelation function, linear predictive coefficients (LPC), and partial correlation (PARCOR) coefficients. In [6] the most efficient representation according to the mentioned above requirements were the PARCOR coefficients. The system which used them acheived 100% sucess in recognition, in the shortest time and with minimal amount of data.

White and Neely [7] tested parametric representations which were extracted by sampling the outputs of 6 and 20 channel filter banks, against the log ratio linear predictive residual (LPR) which is based on the LPC. Their conclusion was that the two representations were approximately equivalent in accuracy.

Hence, we tested the following parametric representations: PARCOR coefficients, LPR, and filter bank implemented by software. It is shown in [4] that the amount of computations required for extraction of each of the parametric representations is approximately the same. However in the calssification phase, the PARCOR coefficients require less computations than the other representations.

Since all three representation were found to yield good recognition results, the PARCOR coefficients were chosen in order to minimize recognition time in the implemented system.

The PARCOR coefficients were extrcated by the autocorrelation method and not by the covaiance method, because only in the first method, the all-pole filter associated with PARCOR coefficients is theoretically stable [8], and this staibility is actually achieved with floating point computations.
Although, for pitch synchronous analysis the covariance method gives more accurate estimates of the speech waveform [9], but this analysis requires an additional amount of computations. In the case of pitch asynchrounous analysis with a large segment (2 or 3 times pitch period) the performance of both methods in representing speech waveform is more or less the same [9]. Therefore, the analysis segment size in the implemented system is N=300 samples (30msec for sampling rate of 10KHz). Choosing this value is suported by the recomendation in [10], that $N=\delta F_s$ where $\delta=20$ to 35 and F_s is the sampling rate in KHz.

Prior to the analysis, the speech samples of the compressed utterance were differenced as a simple mean for frequency preemphasis, which is needed for acheiving better a approximation, with the PARCOR coefficients of the vocal tract transfer function [11]. Each differenced analysis segment is multiplied by an Hamming window, having $N_w=N=300$ samples. The algorithm presented by Markel and Gray in [12] for computing the PARCOR coefficients is then applied.

Since the amount of computations required for the PARCOR extraction is proportional to the number of coefficients P, and this step provides in the implemented system most of the computation load, it is desired to use the lowest value of P which still does not degrade recognition performance. Based on the work in [6] and [13] it was decided to use P=6. It is quite reasonable that 6 PARCOR's represent the first 3 formants. The higher formants were attenuated and filtered since the speech signal was band limited to 200-3200 Hz. Hence, as was also verified by us, the accuracey of the representation of the speech signal with six PARCOR's is sufficient .
In order to efficiently approximate the fast changes in the speech signal a set of 6 PARCOR's was computed for every $N_k=150$ samples of the compressed utterance. With $L_{co}=3000$ samples of the compressed utterance, and the above chosen N_k and N_w, $L_s=19$ sets

The first ten Hebrew digits have constituted the vocabulary of words tested. All, but one, of vocabulary words have two syllables and their duration (as measured for three speakers) were in the range of 0.5 to 0.9 seconds.

Two different data bases have been used in the experiments. One data base contained the utterances of three speakers which were recorded in a quite room on an Akai cassette recorder. Each speaker has recorded four sets of the ten digits. The recordings were made in two sessions, several days apart. The second data base contained the utterances of two speakers (out of the previous three),which were recorded, with a regular omni-directional microphone, in the noisy computer room where the experiments were conducted. The signal to noise ratio was found to be in the range of 18 to 21 dB in the frequency band of 200 to 3200Hz used. Again four sets of the ten digits were recorded for each speaker in two separate sessions, several days apart .

In each of the recognition tests, a word in one library (set of ten digits) of a given speaker was compared (using PARCOR representation) to each of the ten words of another library of the same speaker. This way 120 recognition tests were conducted for each speaker in each of data bases. Thus the overall number of tests performed was 360, for the first data base (quiet environment), and 240 for the "noisy data base". The results of the recognition tests for IWRS with TSC and for IWRS with DP are summarized in Tables I and II, respectively.

TABLE I

Summary of Recognition Test Results for IWRS with TSC

Type of Environment	Number of Tests	Number of Errors per Speaker IE	DM	DA	Recognition Score (%)
Quiet	360	3	0	0	99.1
Noisy (S/N=20dB	240	3	2	-	97.9

TABLE II

Summary of Recognition Test Results for IWRS with DP

Type of Environment	Number of Tests	Number of Errors per Speaker IE	DM	DA	Recognition Score (%)
Quiet	360	6	0	0	98.3
Noisy (S/N=20dB)	240	5	0	-	97.9

VII. Discussion and Conclusions.

Comparing the results summarized in Tables I and II, we observe no loss of overall performance of the system with TSC (for the quiet environment, the results with compression were even better), while the overall computational load was reduced by a factor of at least three, as it is shown in Table III.

TABLE III

Amount of Computations in the TSC and DP Systems.

	TSC		DP	
	additi-ons	multipl-ications	addit-ions	multiplica-tion
Detection of beak and end of utterance	42000	-	42000	-
Compression	6000	3000	-	-
Differencing	5700	-	14200	-
PARCOR's extraction	39900	45600	99300	113500
Classification	9180	-	249900	11900
Total	102780	48600	405400	125400

It can be observed from Table III that even if time duration normalization at the classification phase, in a system without TSC, is performed by linear time scaling [7] which requires considerably less computations than DP, the computation load in the system with TSC is still reduced by a factor of at least two.

It is realized that the number of tests performed, the size of vocabulary used, and the number of speakers involved in the above experiments, are not sufficient for reaching at final conclusions with respect to the application of the proposed TSC algorithm to any general all-digital IWRS. Yet, it is believed that the results obtained support the proposed approach and point out to its potential.

It is stated in [7] that severe reduction of the raw data by preprocessing could have an adverse effect on recognition results.
It is conjectured here that the reason that the compression performed did not cause a loss in performance, is due to the way the proposed TSC algorithm performs the compression,namely, weighting of all or most of the input data and not just discarding portions of it.

References

1) D.R. Reddy,"Speech Recognition by Machine:A Review," PROC. IEEE, vol.64, pp.501-531, April 1976.
2) D.Malah, "Time Domain Algorithms for Time Scale Variation of Speech Signals,"Technion-IIT,EE Pub. No. 280, May 1976.
3) D.Malah and I.Engel,"Computation Reduction in All-Digital Isolated-Word Recognition Systems By Efficient Pre-Extraction Time normalization". Technion-IIT, EE Pub. No. 292, Nov. 1976.
4) I.Engel,"A Minicomputer Implementation of a Speech Recognition System",MSc. Dissertation,submitted to the Technion-IIT,Haifa,Israel,January 1977.
5) J.L.Flanagan and R.M.Golden,"Phase Vocoder",Bell Sys. Tech.J., vol. 45,pp.1493-1509, Nov. 1966.
6) A.Ichikawa,Y.Nakano,and K.Nakata,"Evaluation of Various Parameter Sets in Spoken Digits Recognition" IEEE Trans. Audio Electroacoust.,vol.AU-21, pp. 202-209, June 1973.
7) G.M. White and R.B.Neely,"Speech Recognition Experiments with Linear Predicition, Bandpass Filtering, and Dynamic Programming" IEEE Trans.Acoust.Speech, Signal Processing,Vol.ASSP-24,pp.183-188,April 1976.
8) J.Makhoul,"Linear Prediction: A Tutorial Review," Proc. IEEE,vol. 63,No.4,pp.561-580, April 1975.
9) S.Chandra and W.C.Lin,"Experimental Comparison Between Stationary and Nonstationary Formulations of Linear Prediction Applied to Voiced Speech Analysis", IEEE Trans.Acoust.Speech,Signal processing,Vol.ASSP-22,pp.403-415, Dec.1974.
10) J.D.Markel,"Digital Inverse Filtering-A New Tool for Formant Trajectory Estimation",IEEE Trans. Audio Electroacoust. Vol.AU-20,pp.129-137,June 1972.
11) J.D.Markel and A.H. Gray,Jr.,Linear Predicion of Speech,Berlin Heidelberg New York:Springer-Verlag, 1976.
12) ——, "On Autocorrelation Equations as Apllied to Speech Analysis",IEEE Trans. Audio Electroacoust. Vol. AU-21,pp.66-79, April 1973.
13) F.Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition",IEEE Trans.Acoust. Speech,Signal Processing,Vol.ASSP-23,pp.67-72, Feb. 1975.
14) V.M. Velichko and N.G.Zagoruyko,"Automatic Recognition of 200 words",Int. J. Man-Machine Studies, Vol. 2, pp. 223-234, 1970.