PERFORMANCE OF ISOLATED WORD RECOGNITION SYSTEM FOR DEGRADED SPEECH

B. Yegnanarayana
Department of Computer Science
and Engineering,
Indian Institute of Technology
Madras - 600036
INDIA

Sarat Chandran
Department of Computer Science
Yale University
New Haven
Connecticut - 06520
U.S.A.

ABSTRACT

The performance of an isolated word speech recognition (IWSR) system is known to drop rapidly with increase in the degradation of the input speech. In this paper we propose a recognition scheme which adapts itself to mild degradations in speech. The scheme does not need apriori information regarding the nature and extent of noise. We suggest techniques which adaptively discriminate between noisy and noise-free parameters by using a selective weighting procedure in the final distance calculations. A suitable index is used to study the performance of the recognition system for small data sets. Our scheme lends itself to greater flexibility in handling degradations in speech input than do the existing recognition schemes. We illustrate our scheme by simulating an adaptive differential pulse code modulated (ADPCM) speech, where the main distortion is contributed by the quatization noise.

I. INTRODUCTION

This paper describes a scheme to improve the performance of an Isolated Word Speech Recognition (IWSR) system in the presence of degradation in the input speech. We make use of the concept of signal-dependent matching [1]. The basic system performs a spectral match using a dynamic time warping (DTW) algorithm [2]. The philosophy behind our scheme is to identify the high SNR regions of the input spectrum and increase their contribution towards the final calculated distance. We study three techniques for achieving this discrimination. All these techniques depend on determining a weighting function for the parameters derived from the input test data. The stored templates themselves are not affected in the process. To reduce the computational load, and to permit verification of the performance of the schemes on different data sets, only small (5 words) vocabularies are considered. The relative performance of various techniques are assessed using a performance index [1] based on the distance matrix. This, in our opinion, gives a better indication of performance for small data sets, than does the normal statistical performance index given by the percentage of correct recognition.

The paper is organized as follows. In Section II we describe the proposed signal-dependent matching for degraded speech input. We describe our recognition experiments in Section III and the results of our recognition experiments in Section IV.

II. SIGNAL-DEPENDENT MATCHING FOR DEGRADED SPEECH

2.1 Basic IWSR System

Presently most IWSR systems yield a recognition accuracy of more than 95% . This figure, however, is dependent on various factors like the size and nature of the vocabulary, the reliability and robustness of the parameters, background noise, and so on [3]. We are concerned with the deterioration in performance caused by degradation in the input speech. To investigate this aspect of the system performance, we select a standard IWSR system in which other design parameters are held constant in all experiments.

The parameters chosen to represent the speech are the log melspectral parameters [4]. Let $M_r(k)$ and $M_t(k)$, k=1, 2....16, be the 16 parameters of the reference and test frames respectively. We define the frame distance 'd' as

$$d = \sum_{k=1}^{16} | M_r(k) - M_t(k) | \qquad (1)$$

The overall distance (D) is evaluated from the frame distances using a DTW algorithm [2].

2.2 Overview of Signal-dependent Matching

In attempting to develop a system which adapts itself to degradation in the input speech, we introduce the concept of signal-dependent matching (SDM) [1]. SDM aims at a more efficient use of the information content of the signal. By this

CH1945-5/84/0000-0094 $1.00 © 1984 IEEE

method, a discrimination is first achieved between the information bearing parameters and the less important ones. Weightages are then given to various parameters so that the overall parameter set will contribute more effectively towards the ultimate computed distance between two words.

We define a modified frame distance:

$$d = \sum_{k=1}^{16} W(k) \mid M_r(k) - M_t(k) \mid, \qquad (2)$$

where $W(k)$ is the weight function. The mode of selection of this function is described in the following sections.

2.3 Choice of Weight Function

In implementing the SDM techniques in a recognition system one must first have an idea of the manner in which external degradation affects the speech spectra. The weight function is then applied so as to emphasize the contribution of the high SNR (less degraded) regions towards the distance, compared to the contribution of the lower SNR (highly degraded) regions. It is important to note here that the weight function should be derived from the test utterance frame and not from the reference.

We propose three methods for deriving the weight function from the spectral parameters of a test frame.

Method-A : Negative derivative of phase spectrum (NDPS)
Method-B : Normalised Zero mean spectrum
Method-C : Peak detection

Method-A : Negative Derivative of Phase Spectrum

It is known [5] that the Negative Derivative of Minimum Phase Spectrum (NDPS) of a smoothed log spectrum can be effectively used to isolate regions of peaks and valleys of the spectrum. This seems to fit in with our requirement of a method of discriminating between high and low SNR regions.

Let $\Theta'(k)$, $k = 1,2....16$ be the 16 NDPS values derived from a particular time frame. We define the weighting function $W(k)$.

$$W(k) = 1, \quad \Theta'(k) > 0 \qquad (3)$$

$$= \epsilon, \quad \Theta'(k) \leq 0,$$

where $0 \leq \epsilon \leq 1$.

Method-B : Normalized Zero Mean Spectrum

A drawback of the NDPS technique is that it makes inadequate use of the knowledge of the frequency domain behaviour of speech. In a typical voiced speech spectrum, the low frequency regions have a prominently higher energy content. In Method-B we exploit this property as follows: The spectrum is normalised to zero mean by computing the deviation of each spectral value from the mean of the 16 spectral values of a frame. The negative values are weighted by a factor ϵ. The choice of optimum ϵ is to be made experimentally.

Method-C: Peak Detection

Method-B does not permit sufficient control over the contribution of parameters in different frequency regions. Since the main objective in the development of a weighting function is to distinguish between the peak and valley regions, we use a direct procedure to detect and weight the peaks. Each spectral point is compared with its immediately adjacent neighbours. If it is greater than both, then that spectral value is given a weightage unity. The immediate neighbours are given a weightage ϵ_1, while all other points get a lower weightage ϵ_2.

III. RECOGNITION EXPERIMENTS

3.1 Data Preparation

Speech data was selected from two repetitions (by a female speaker) of 36 words of the alpha-digit task (the alphabet A-Z and the digits 0-9). For our investigations we generated degraded speech from the given speech data. Our choice of the type of degradation was dictated by the following considerations:

(a) It should be naturally occuring

(b) It should be possible to vary the level of degradation conveniently.

The quantization noise in ADPCM coded speech is chosen for our studies. By varying the number of quantization levels (bits) we can control the extent of degradation. The ADPCM coded speech is generated using the scheme described in [6]. We have also investigated the effect of additive random noise by adding random noise to the sampled speech at a predetermined signal to noise ratio (SNR)level.

One repetition of the words is used to create the reference templates and the other repetition is used as the test utterances to be recognised. Each 25.6 msec segment of speech is considered as a frame and is represented by 16 log melspectral values. The NDPS values for each frame of the test data are computed from the log melspectral coefficients using a 32 point DFT [5].

Recognition tests are carried out on

small vocabularies of 5 words each. The IWSR system compares each test utterance with the templates of each reference word and produces a distance matrix. We employ the performance index (PIX) developed in [1] to describe a distance matrix.

3.2 Experiments

Experiment-1 :

The first set of experiments was designed to determine the optimum value of the weight parameter ϵ for Method-A and B and ϵ_1 and ϵ_2 for Method-C.

Experiment-2 : (Results in Table-1)

The performances of the system were examined for the cases of direct matching (no SDM) and the SDM methods A, B and C. This was done by fixing the value of ϵ, based on the results of Experiment-1, in each scheme. The performance was evaluated for four vocabularies V_1, V_2, V_3, V_4 for both ideal and degraded (2-bit ADPCM) speech inputs.

Experiment-3 : (Results in Table-2)

The performances of the 4 schemes (direct matching and three SDM methods) for varying conditions of noise were studied. The vocabulary V_1 was chosen for this study and the value of ϵ is same as that in Experiment-2. The study was conducted for three different degradations in the input speech, namely, 2-bit ADPCM, 3-bit ADPCM and additive pseudo-random noise.

IV. RESULTS AND DISCUSSIONS

The results of Experiment-1 showed a significant improvement in the recognition of degraded speech after incorporation of the SDM techniques. We also noticed that the PIX in each method was not greatly influenced by small variations in the deemphasizing weight ϵ. Therefore the value of ϵ is not very critical as long as it is in the range 0.1 to 0.4.

From Table-1, we see that the value of PIX changes for different vocabularies depending on the inherent confusability of the vocabularies. In all cases SDM gives a substantial improvement over the conventional matching technique. An important criterion in evaluating the suitability of a particular technique lies in ensuring that while the performance may improve for the degraded speech input, the performance should not drop (significantly) for the normal input speech. By this criterion Method-C (Peak Detection), which gives an improved performance for both normal and degraded speech, seems to be the best choice.

For the set V_4 (A,B,F,I,Z), the overall performance is poor and the improvement due to SDM is not very significant. In fact, the use of the NDPS method has led to a lower PIX values. This is because the peaks and valleys in the spectral characteristics of the fricatives 'F' and 'Z' do not have the same significance as the characteristic resonances of voiced speech. Emphasis of peaks in the SDM process in such cases may not be very meaningful.

Table-2 indicates the general improvement in performance due to SDM for different noise environments. Here also the peak detection method gives consistently large improvement for all the three types of noise. The percentage improvement in performance with SDM is far greater for the 2-bit ADPCM than for the 3-bit ADPCM speech. Further, since the general noise level is lower for the 3-bit ADPCM case, the peak detection method gives a superior PIX value compared to the zero-mean method.

We believe that the performance measure PIX gives a better representation of the system performance than the statistical measure of percentage of correct recognitions. We have observed that the percentage recognition score has not dropped after incorporation of the SDM.

V. SUMMARY AND CONCLUSIONS

In this paper we presented methods of improving the performance of a speech recognition system for degraded speech input by using a signal-dependent matching (SDM) strategy. We apply the SDM to emphasize the contribution of important features towards the computation of the distance between two words. The peaks in the spectra are seen to be clearly less degraded and are hence used to contribute more towards the frame to frame distance through the generation of a weight function.

In our study we have confined ourselves to a restricted data set as well as to a limited variation in the type of degradation. Experiments should be carried out over a wider range of vocabularies to study the full implication of the SDM. We have also not addressed the computational issues of the proposed schemes. All the three methods involve many time consuming steps in the computation of the individual frame distances, which will lead to a marked increase in the response time of the recognition system.

REFERENCES

[1] B.Yegnanarayana and T.Sreekumar, 'Signal Dependent Matching for Isolated Word Speech Recognition Systems', (To be published in Signal Processing)

[2] H.Sakoe and S.Chiba, 'Dynamic Programming Algorithm Optimization for Spoken Word Speech Recognition Systems', IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-28, Feb. 1978, pp. 43-49.

[3] L.R.Rabiner and S.E.Levinson, 'Isolated and connected word recognition - Theory and selected applications', IEEE Trans. Communications, Vol.COM-29, May 1981, pp. 621-659.

[4] Steven B.Davis and Paul Mermelstein, 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', IEEE Trans. Acoustics, Speech, Signal Processing, Vol.ASSP-28, Aug. 1980, pp. 357-366.

[5] B. Yegnanarayana, 'Pole-Zero Decomposition of speech spectra', Computer Sciences Dept., CMU, Pittsburgh, PA-15213, Report : CMU-CS-79-101, Jan. 1979.

[6] M.R.Sambur and N.S.Jayant, 'LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise', IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-24, Dec.1976, pp.488-494.

Table-1 : Comparison of Schemes on Different Vocabularies

V_1 : /O,1,2,3,4/
V_2 : /5,6,7,8,9/
V_3 : /L,O,P,Q,R/
V_4 : /A,B,F,I,Z/
Reference Data : Normal Speech
Test data : 2-bit ADPCM speech

Vocabulary	Test data type	Conventional (No SDM)	SDM Method-A ($\epsilon = 0.2$)	SDM Method-B ($\epsilon = 0.2$)	SDM Method-C ($\epsilon_2 = 0.2$)
V_1	Degraded	85.1	89.7	93.5	90.4
V_2	,,	95.5	98.1	98.6	97.8
V_3	,,	80.1	95.7	95.7	87.2
V_4	,,	90.0	89.0	91.3	91.7
V_1	Normal	98.3	98.8	97.9	99.9
V_2	,,	99.7	99.9	99.8	99.9
V_3	,,	89.0	98.4	99.4	96.6
V_4	,,	92.9	91.6	91.5	93.4

Table-2 : Comparison of schemes for different noise environments

Reference data : Normal Speech
Test data : Degraded Speech
Vocabulary : V_1

Type of Degradation	Conventional (No SDM)	SDM Method-A ($\epsilon = 0.2$)	SDM Method-B ($\epsilon = 0.2$)	SDM Method-C ($\epsilon_2 = 0.2$)
2-bit ADPCM	85.1	89.7	93.5	90.4
3-bit ADPCM	95.9	96.1	96.7	97.8
Additive Noise	85.0	87.7	84.3	96.4