Signal Processing 7 (1984) 175-183
North-Holland

ON IMPROVEMENT OF PERFORMANCE OF ISOLATED WORD RECOGNITION FOR DEGRADED SPEECH

B. YEGNANARAYANA
Dept. of Computer Science and Engineering, Indian Institute of Technology, Madras-600036, India

SARAT CHANDRAN
Dept. of Computer Science, Yale University, New Haven, Connecticut, CT 06520, USA

ANANT AGARWAL

Received 22 September 1983
Revised 7 December 1983 and 5 April 1984

Abstract. We investigate the performance of an isolated word speech recognition (IWSR) system for degraded speech. We propose a recognition scheme which adapts itself to mild degradations in speech and improves the reliability of recognition significantly. The scheme does not use a priori 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 calculation. A new measure of performance is adopted to compare several recognition schemes using small data sets. Our scheme lends itself to greater flexibility in handling degradations in speech input than do the existing recognition schemes.

Zusammenfassung. Wir untersuchen die Wirkungsweise eines Systems zur Erkennung isoliert gesprochener Wörter für Sprachsignale verminderter Qualität. Vorgeschlagen wird ein Erkennungschema, das sich bei geringer Qualitätsverminderung an die Qualität des Signals anpaßt und die Gesamterkennungsrate spürbar verbessert. Das Schema benötigt kein A-Priori-Wissen über die Art und Stärke des Hintergrundrauschens. Die vorgeschlagenen Methoden unterscheiden adaptiv zwischen verrauschten und rauschfreien Parametern, indem sie eine selektive Gewichtungsfunktion bei der Berechnung der Abstandsmaße einsetzen. Um verschiedene Worterkennungssysteme mit kleinen Datenmengen vergleichend zu testen, wird ein spezielles Maß für die Erkennungssicherheit eingeführt. Mit Hilfe unserer Methode lassen sich Störungen und Qualitätsverminderungen bei der Spracheingabe leichter und flexibler behandeln als mit herkömmlichen Spracherkennungssystemen.

Résumé. Nous nous intéressons aux performances d'un système de reconnaissance de mots isolés, en présence de parole dégradée. Nous proposons un schéma de reconnaissance qui s'adapte aux légères dégradations de la parole, et améliore sensiblement la fiabilité de la reconnaissance. Ce schéma n'utilise aucune information a priori sur la nature et la quantité du bruit. Nous suggérons des techniques qui discriminent de façon adaptative les paramètres bruités et les non-bruités, au moyen d'une procédure de pondération sélective de la distance finale. Une nouvelle mesure de qualité est employée pour comparer divers schémas de reconnaissance sur de petits ensembles de test. Notre procédure offre une plus grande souplesse pour prendre en compte les dégradations du signal d'entrée que les schémas de reconnaissance existants.

Keywords. performance, isolated word recognition, selective weighting procedure.

1. Introduction

Most practical speech recognition systems are required to function in environments that are not entirely noise-free. The performance of an Isolated Word Speech Recognition (ISWR) system is known to drop rapidly with increase in the degradation of the input speech. Methods such as noise normalization, template tuning, etc., have been proposed [1], by assuming a priori noise statistics. Such systems are designed to function optimally for a particular noise environment. A change in the degrading noise characteristics will lead to suboptimal performance, with a consequent drop in the recognition accuracy.

This paper describes a scheme to improve the performance in the presence of degradation. We make use of the concept of signal-dependent matching [2]. The basic system performs a spectral match using a dynamic time warping (DTW) algorithm [3]. 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 word) vocabularies are considered. The relative performance of various techniques are assessed using a performance index described in [2].

The paper is organized as follows. Section 2 describes the nature of the degradation problem in IWSR systems. In Section 3 we develop three procedures to tackle this problem via signal-dependent matching. Section 4 describes the experimental details and the methodology of our investigation. Section 5 is devoted to a discussion of the results of our experimental studies, and Section 6 summarizes the paper and discusses some of the limitations of our approach.

2. 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 [4]. 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 magnitude melspectral parameters [5]. Let $M_r(k)$ and $M_t(k)$, $k = 1, 2, \ldots, 16$, be the 16 parameters of the reference and test frames respectively. We define the frame distance 'd' as

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

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

In this paper we confine ourselves to the category of noise broadly classified as additive noise. The performance of the system is critically dependent on the susceptibility of the parameters to noise. Fig. 1 shows the spectrum of a speech segment

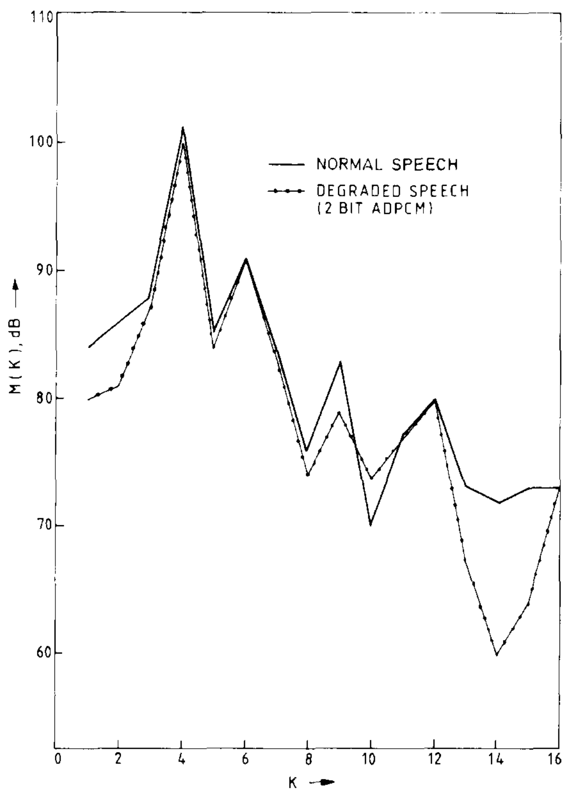

Fig. 1. Degradation of speech spectrum caused by 2-bit ADPCM coding.

corrupted by ADPCM coding noise [6]. The use of the corrupted parameters will naturally decrease the accuracy and reliability of recognition.

3. Signal dependent matching for degraded speech

In attempting to develop a system which adapts itself to degradation in the input speech, we introduce the concept of signal-dependent matching (SDM) [2].

3.1. Overview of SDM

SDM aims at a more efficient use of the information content of the signal. By this 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

$$d = \sum_{k=1}^{16} W(k)\left|M_r(k) - M_t(k)\right|, \tag{2}$$

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

3.2. Choice of weight function

In implementing the SDM techniques in a recognition system, one must first have an idea of the manner in which the external degradation affects the speech spectra. The weight function is then applied so as to emphasise 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. The rationale behind this is as follows. Each test utterance is compared with each of the reference words and a decision is reached regarding its identity based upon the minimum distance criterion. If the weights are derived from the reference templates, comparison of the test utterance with various references would yield distances which depend on the weight distribution in the template, which will be different for different reference words. Moreover, the weights derived from the reference words are completely independent of the degradation in the input speech.

The spectral envelope of a voiced speech segment is characterised by the presence of distinct peaks, corresponding to the resonances of the vocal tract system. It is logical to assume that these spectral peaks are more important for the final recognition than the spectral valleys—a fact that is substantiated by the success of all-pole models for such segments. Further, additive noise is more likely to affect the lower signal power (valley) regions of the spectrum. The marked degradation of the valleys relative to the peaks is noticed in Fig. 1. Our choice of the weight function will therefore aim at increasing the bias of the peak regions of the short-time spectral envelope.

3.3. Weight function generation

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: Normalized zero mean spectrum

Method C: Peak detection

3.3.1. Method A: Negative derivative of phase spectrum

It is known [7], [8] that the Negative Derivative of minimum Phase Spectrum (NDPS) of a smoothed log spectrum can be effectively used to isolate the 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. The characteristics of the NDPS can be seen in Fig. 2, where a spectrum and its NDPS are shown. The regions of positive NDPS correspond to the peak regions and the regions of negative NDPS indicate the valley regions of the spectrum.

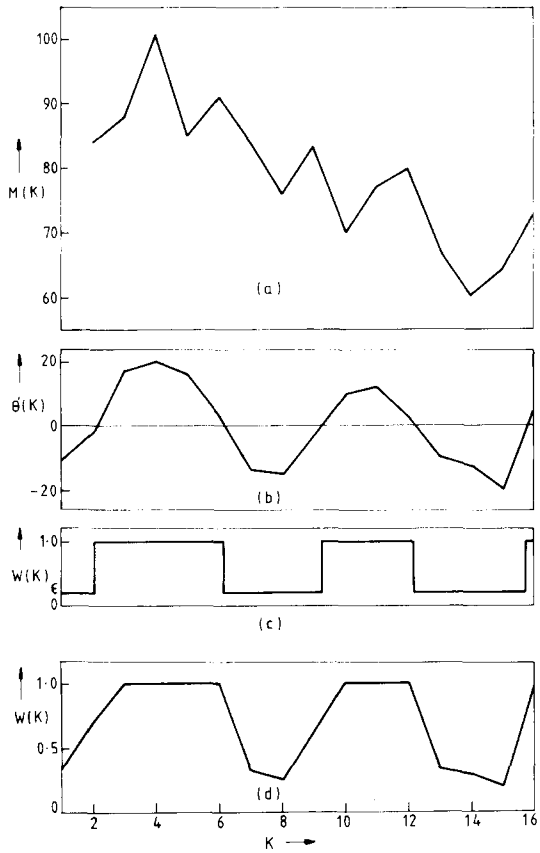

Fig. 2. Weight function generation using the NDPS method (Method A). (a) The 16 log melspectral parameters, (b) NDPS of spectrum, (c) Constant ε weight function, (d) Variable ε weight function.

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

$$W(k) = 1, \quad \theta'(k) > 0$$
$$= \varepsilon, \quad \theta'(k) \leq 0, \tag{3}$$

where $0 \leq \varepsilon \leq 1$.

We feel that the optimum ε must lie between 0 and 1, because setting $\varepsilon = 1$ is tantamount to direct matching ($W(k) = 1$ for all k in eq. (2)). On the other hand, setting $\varepsilon = 0$ removes the valley regions of the spectrum from consideration, but leads to loss of some signal information also. Poorer recognition will therefore result in both these extreme cases. There is no theoretical basis on which we can select the optimum ε. In fact ε could even be defined as a function of frequency, so that the deemphasis is not abrupt, but goes through a gradual transition about the NDPS zero crossing. We have chosen a variable function of the form

$$W(k) = 1, \qquad \theta'(k) > 0$$
$$= \varepsilon = \frac{-p}{(-p + \theta'(k))}, \quad \theta'(k) \leq 0, \tag{4}$$

where p is a positive integer used to set the range of ε. Also note that $\theta'(k)$ being negative, ε can only assume positive values. Examples of constant and variable ε functions are shown in Fig. 2c and Fig. 2d.

3.3.2. 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. The weight function described above gives equal importance to spectral values over the entire frequency range. That is, a valley in the high frequency regions is deemphasised only as much as the one in the lower frequencies. However, we know that, for typical voiced speech segments, the major portion of the signal informaton is contained in the first and second formants, while the contribution of the remaining spectral peaks is not as significant. We also know that, in general, additive noise has greater spectral power in the higher frequencies. This fact can be observed in the spectral plot in Fig. 1—where we see that the degradation is far greater in the valleys situated at higher frequencies. Based on the above observations, we develop a technique which automatically gives lesser weightage to the higher frequency regions.

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 (Fig. 3a).

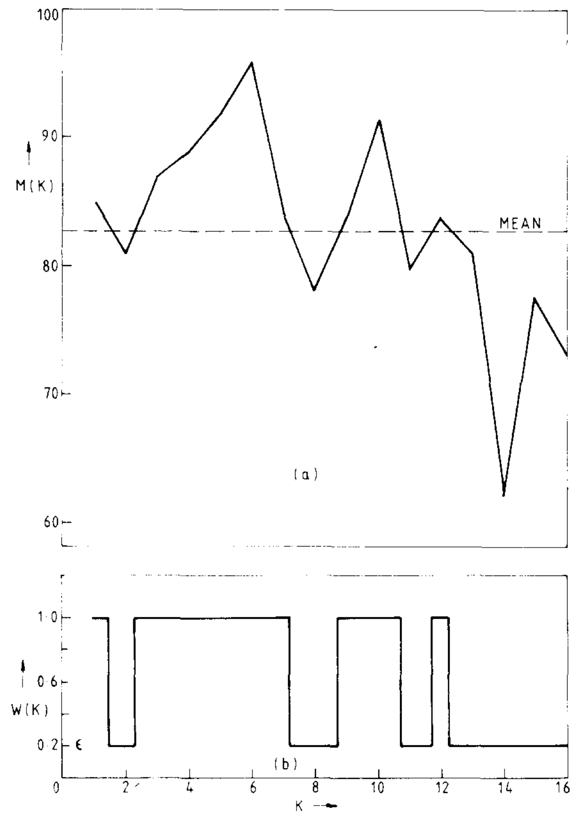

Fig. 3. Weight function generation using the zero mean normalization method (Method B). (a) Speech spectrum, (b) Constant ε ($\varepsilon = 0.2$) weight.

Let $M_n(k)$, $k = 1, 2, \ldots, 16$ be the normalised spectral values of a frame of test utterance. The weight function (Fig. 3b) is distinctly biased towards the low frequency regions, as desired. Here too, ε can be a constant or a variable and the choice of optimum ε is to be made experimentally.

3.3.3. Method C: Peak detection

Method B has the following shortcomings: (1) The high frequency peaks do carry some information useful for recognition. Through the zero-mean normalisation process, this information may be lost, as can be seen for the case of the fourth peak in Fig. 3a. (2) While some discrimination should exist between the weightage given to the peaks in different frequency ranges, we are not justified in giving a weightage of 1.0 to the first peak region and a near zero weightage to the last one (Fig. 3b). Method B, therefore, 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 1.0 weightage. The immediate neighbours are given a weightage ε_1, while all other points get a lower weightage ε_2. An example of a weight function derived from the peak detection procedure is shown in Fig. 4. The values of ε_1 and ε_2 are decided by experimentally observing the performance of the IWSR system. The high frequency regions can now be conveniently de-emphasised as desired, by multiplying the weight

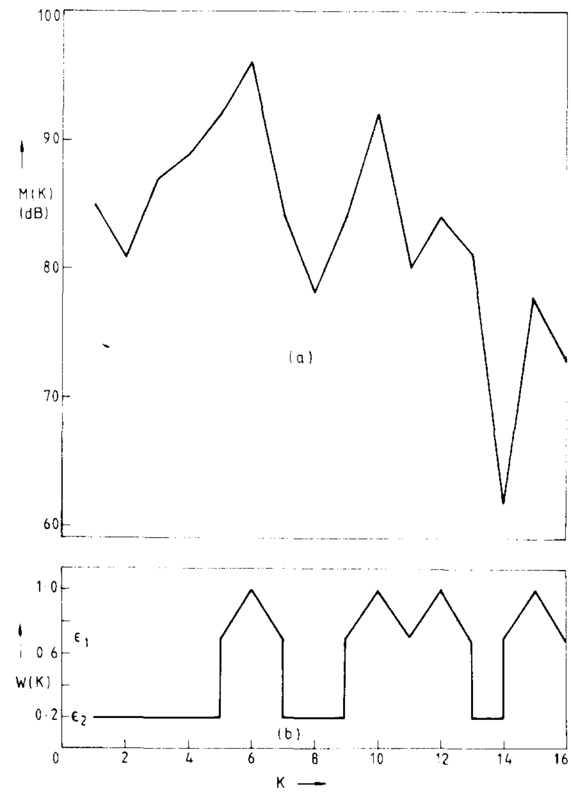

Fig. 4. Weight function generation using peak detection (Method C). (a) Speech spectrum, (b) Weight function with $\varepsilon_1 = 0.7$, $\varepsilon_2 = 0.2$.

function $W(k)$ with a suitable deemphasis function $D(k)$. That is

$$W'(k) = W(k) \cdot D(k), \quad k = 1, 2, \ldots, 16, \tag{5}$$

4. Recognition experiments

4.1. Data preparation

Speech data was selected from two repetitions (by a female speaker) of 36 words of the alphadigit 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:

(i) It should be naturally occurring.

(ii) 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 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 recognized. 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. The procedure for computing the NDPS is described in [7].

Recognition tests were carried out on small vocabularies of 5 words each. The choice of the size of the vocabulary was dictated by the available facilities for computation. We find that even small vocabularies are adequate to illustrate the significance of signal-dependent matching for degraded speech. Four different subsets of vocabulary are used in our studies. They are:

V_1 – 'Zero', 'One', 'Two', 'Three', 'Four'

V_2 – 'Five', 'Six', 'Seven', 'Eight', 'Nine'

V_3 – 'L', 'O', 'P', 'Q', 'R'

V_4 – 'A', 'B', 'F', 'I', 'Z'

The IWSR system compares each test utterance with the template of each reference word and produces a distance matrix. We employ the performance index (PI) proposed in [2] to evaluate the performance of a recognition system under different conditions of operation.

4.2. Experiments

Experiment 1: (Results in Tables 1a, 1b, 1c)

The first set of experiments was designed to determine the optimum value of the weight parameter ε for Methods A and B and ε_1 and ε_2 for Method C. For this study we used the vocabulary V_1 and the 2-bit ADPCM speech as the degraded test input. A variable weight function as given by eq. (4) was also studied for Methods A and B.

Experiment 2: (Results in Table 2)

The performance of the system was 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 for each scheme. The performance was evaluated for the four vocabularies V_1, V_2, V_3, V_4 for both normal and degraded (2-bit ADPCM) speech inputs.

Experiment 3: (Results in Table 3)

The performance of the 4 schemes (direct matching and three SDM methods) for varying conditions of noise was 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 pseudorandom noise.

Experiment 4: (Results in Table 4)

This is intended to illustrate a special instance of the recognition problem. Sometimes the templates are created from noisy utterances. If conditions remain the same in the testing situation, a reasonable (improved) level of performance can be anticipated. This is the case of 'template-tuning'. If, however, the noise has vanished from the testing environment, then the drop in performance due to degraded templates will be felt. Here, we use the vocabulary V_1 to test the four schemes. The templates are created using the 2-bit ADPCM speech data and tested with normal and degraded utterances.

5. Results and discussion

In Tables 1a, 1b and 1c, we notice a significant improvement in the recognition of degraded speech after incorporation of the SDM techniques. We also notice that the PI in each method is not greatly influenced by small variations in the de-emphasising weight ε. Therefore the value of ε is not very critical as long as it is in the range 0.1 to 0.4.

For Method A (Table 1a), the performance for the degraded input case is best for $\varepsilon = 0.2$. This is because a low ε causes loss of signal information and a high ε gives too much prominence to noise. It is interesting to note that the performance for normal speech was not affected due to weighting of different spectral components.

With Method B (Table 1b), we again get the best recognition for $\varepsilon = 0.2$. As anticipated, a marked improvement in the recognition of degraded speech is observed. But, the performance for normal speech has deteriorated for $\varepsilon = 0$ and $\varepsilon = 0.2$. This can be explained in terms of the loss of high frequency information without any compensating reduction in noise.

Method C (Table 1c) gives its best performance for $\varepsilon_2 = 0$. The most striking feature of this scheme is the significant improvement in performance for both normal and degraded speech inputs. This

Table 1

Results of Experiments to determine optimum value of the weight parameter ε. Vocabulary: V_1; Reference data: Normal Speech; Degradation: 2-bit ADPCM; PI without SDM ($\varepsilon = 1$) for degraded input: 85.1, for normal input: 98.3

Table 1a: NDPS (Method A)

Test data: Degraded speech		Test data: Normal speech	
ε	PI	ε	PI
0.0	89.1	0.0	98.6
0.2	89.7	0.2	98.8
0.4	88.5	0.4	98.8
$-5/(-5+\theta'(k))$	88.8	$-5/(-5+\theta'(k))$	98.9

Table 1b: Zero-mean normalization (Method B)

Test data: Degraded speech		Test data: Normal speech	
ε	PI	ε	PI
0.0	93.3	0.0	97.5
0.2	93.5	0.2	98.0
0.4	92.7	0.4	99.2
$-5/(-5+M_n(k))$	88.8	$-5/(-5+M_n(k))$	98.7

Table 1c: Peak detection (Method C)

Test data: Degraded speech		Test data: Normal speech	
ε_2 ($\varepsilon_1 = 0.7$)	PI	ε_2 ($\varepsilon_1 = 0.7$)	PI
0.0	91.3	0.0	99.9
0.2	90.4	0.2	99.9
0.4	89.4	0.4	99.6

supports our original assumption that emphasis of peaks in the matching process should aid in the recognition of both normal and corrupted utterances.

From Table 2, we see that the value of PI 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,

Table 2

Comparison of Schemes for different vocabularies. V_1: /0, 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 ($\varepsilon = 0.2$)	SDM Method B ($\varepsilon = 0.2$)	SDM Method C ($\varepsilon_2 = 0.2$)
V_1	Degraded	85.1	89.7	93.5	90.4
V_2	Degraded	95.5	98.1	98.6	97.8
V_3	Degraded	80.1	95.7	95.7	87.2
V_4	Degraded	90.0	89.0	91.3	91.7
V_1	Normal	98.3	98.8	97.9	99.9
V_2	Normal	99.7	99.9	99.8	99.9
V_3	Normal	89.0	98.4	99.4	96.6
V_4	Normal	92.9	91.6	91.5	93.4

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, use of the NDPS method has led to a lower PI. 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 characteristics resonances of a voiced speech. Emphasis of peaks in the SDM process in such cases may not be very meaningful.

Table 3 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 PI to the zero-mean method.

On comparison of the first rows of Tables 2 and 4, we can observe the significant advantages of a priori knowledge of the noise characteristics and template tuning. The results in the first row of Table 4 indicate that the use of SDM leads to an improvement only with Method C. With the reference degraded and the test normal, a drop in PI is noticed on comparison of the 2nd row of Table 4 with the 5th row of Table 2.

We believe that the performance measure PI 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 the incorporation of SDM.

Table 3

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 ($\varepsilon = 0.2$)	SDM Method B ($\varepsilon = 0.2$)	SDM Method C ($\varepsilon_2 = 0.2$)
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

Table 4
Performance of speech recognition system using degraded speech as reference data. Reference data: 2-bit ADPCM speech; Vocabulary: V_1

Test data type	Conventional (No SDM)	SDM Method A ($\varepsilon = 0.2$)	SDM Method B ($\varepsilon = 0.2$)	SDM Method C ($\varepsilon_2 = 0.2$)
Degraded (2-bit ADPCM)	93.6	94.0	93.6	96.6
Normal	96.8	95.0	96.8	97.1

6. Conclusion

In this paper we presented methods of improving the performance of a speech recognition system for a degraded speech input by using a signal dependent matching (SDM) strategy. We have applied 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. The weight function was derived from the test data using (A) NDPS, (B) Zero-mean Spectral Normalisation and (C) Peak Detection techniques. All the three methods yield significant improvements in the recognition performance for degraded speech. The peak detection technique appears to provide the best and most consistent performance among the three methods.

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 of the individual frame distances, which will lead to a marked increase in the response time of the recognition system.

References

[1] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, Vol. 67, No. 12, Dec. 1979, pp. 1586–1604.

[2] B. Yegnanarayana and T. Sreekumar, "Signal dependent matching for isolated word speech recognition systems", *Signal Procesessing*, Vol. 7, No. 2, 1984, pp. 161-173.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-28, Feb. 1978, pp. 43–49.

[4] 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.

[5] S.B. Davis and P. 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.

[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.

[7] B. Yegnanarayana, "Pole-zero decomposition of speech spectra", Computer Sciences Dept., CMU, Pittsburgh, PA-15213, Report: CMU-OS-79-101, Jan. 1979.

[8] B. Yegnanarayana,"Formant extraction from linear perdiction phase spectra", *J. Acoustic. Soc. Amer.*, Vol. 63, 1978, pp. 1638–1640.