Signal Processing 7 (1984) 161-173
North-Holland

SIGNAL-DEPENDENT MATCHING FOR ISOLATED WORD SPEECH RECOGNITION SYSTEMS

B. YEGNANARAYANA and T. SREEKUMAR*

Department of Computer Science and Engineering, Indian Institute of Technology, Madras-600 036, India

Received 28 April 1983
Revised 31 August 1983 and 5 April 1984

Abstract. A new approach to the design of IWSR systems is proposed in this paper. This involves a dynamic matching strategy based on the nature of the input speech segment. This is called signal-dependent matching. The computational complexity in the implementation of the proposed algorithm is significantly reduced by adopting a two stage approach in matching. In the first stage, the warping path between the test utterance and a reference utterance is determined. In the second stage, the distance between the utterances is computed along the path. There will be a slight degradation in the performance of a two stage approach as compared to the single stage approach, but this can be tolerated in view of the significant computational advantage. The performance degradation is more than compensated by the signal-dependent matching strategy in the second stage. To measure the improvement in the recognition performance, a new index of performance is defined, that reflects the characteristics of the distance matrix for a given vocabulary, rather than the characteristics of the confusion matrix. The performance of the signal-dependent matching algorithm is significantly better than the standard dynamic time warping matching algorithm for confusable as well as nonconfusable vocabulary.

We also develop a signal-dependent matching algorithm, which takes into account some distortions in the input speech. As an example we offer the agorithm twice the same test utterance, once undistorted, once after a distortion. Our research until now indicates a improvement in automatic isolated word speech recognition systems while using signal-dependent parameter measuring and signal dependent matching.

Zusammenfassung. In diesem Beitrag wird eine neue Methode zum Entwurf von Systemen zur automatischen Erkennung isoliert gesprochener Wörter vorgeschlagen. Hierbei wird eine Strategie des dynamischen Mustervergleichs eingesetzt, die auf den Eigenschaften des am Eingang anliegenden Sprachsegments beruht; diese Strategie wird als "signalabhängiger Mustervergleich" bezeichnet. Die Komplexität des Algorithmus wird beträchtlich reduziert, wenn der Mustervergleich in zwei Stufen durchgeführt wird. In der ersten Verarbeitungsstufe wird der optimale Pfad für den Mustervergleich zwischen dem Testmuster und einem Referenzmuster ermittelt. In der zweiten Stufe wird dann der Abstand zwischen den beiden Äußerungen unter Zugrundelegung dieses Pfades berechnet. Verglichen mit dem einstufigen Verfahren arbeitet das zweistufige geringfügig schlechter; dies wird jedoch ausgeglichen durch den erheblich verminderten Rechenaufwand der zweistufigen Lösung. Im übrigen wird dieser Mangel bei weitem durch den in der zweiten Stufe eingesetzten signalabhängigen Mustervergleich kompensiert. Um die Verbesserung in der Wirkungsweise quantitativ zu messen, wird ein neues Maß für die Erkennungssicherheit eingeführt, welches für ein gegebenes Vokabular die Eigenschaften der Abstandsmatrix und nicht die der Verwechslungsmatrix berücksichtigt. Die Strategie des signalabhängigen Mustervergleichs arbeitet erheblich besser als das Sttandardverfahren der dynamischen Verzerrung der Zeitachse; dies gilt für Vokabulare, die von Haus aus für Verwechslungen anfällig sind, als auch für solche, die von der Auswahl der Wörter her Verwechslungen wenig wahrscheinlich machen.

In einer Erweiterung wird das Prinzip des signalabhängigen Mustervergleichs auch dazu benützt, auch im Testmuster einige Verzerrungen zu berücksichtigen. Dadurch, daß für jeden Parametermeßpunkt im Testmuster die Parameter adaptiv ausgesucht werden, kann das Verfahren als Ganzes erheblich verbessert werden. Als Beispiel werden die gleichen Testmuster dem Algorithmus einmal unverzerrt, zum anderen nach Verzerrung durch adaptive Differenz-Pulz-Code-Modulation (ADPCM) angeboten. Wie unsere vorläufigen Untersuchungen zeigen, tragen signalabhängige Parametermessung und signalabhängiger Mustervergleich erheblich zur Verbesserung der Erkennungsrate und Sicherheit von Systemen zur automatischen Erkennung isoliert gesprochener Wörter bei.

* Present address: T. Sreekumar, Dept. of Computer Science, Case Western Reserve Univ., Cleveland, OHIO-44106, U.S.A.

Résumé. On propose dans ce papier une nouvelle approache pour la conception des systèmes de reconnaissance de mots isolés. Cette approche utilise une stratégie de comparaison dynamique basée sur la nature de la parole en entrée. C'est ce que nous appelons une comparaison dépendant du signal. La complexité de calcul de l'algorithme proposé est réduite de façon significative par l'emploi d'une comparaison en deux étapes. Durant la première étape, on détermine le chemin de coincidence entre la séquence de test et une séquence de référence. Au cours de la seconde étape, la distance entre les deux séquences est évaluée le long de ce chemin. Il y aura une légère dégradation des performances avec une approche à deux niveaux, par comparaison à une approche à un seul niveau, mais ceci peut être accepté, au vu de l'avantage significatif en complexité de calcul. La dégradation de performances est plus que compensée par la stratégie de comparaison adaptée au signal, qui est employée dans la seconde étape. Pour mesurer l'amélioration de la qualité de la reconnaissance, on définit un nouvel indice de performance, qui reflète plus les caractéristiques de la matrice des distances, pour un vocabulaire donné, que la matrice de confusion entre mots. Les performances de l'algorithme de comparaison dépendant du signal sont sensiblement meilleures que celles obtenues avec un algorithme standard de comparaison dynamique, aussi bien avec des vocabulaires ambigus qu'avec des vocabulaires non ambigus.

Le concept de comparaison dépendant du signal est étendu pour également prendre en compte certaines distorsions de la parole. En choisissant de façon adaptative les paramètres pour chaque échantillon de la séquence de test, nous montrons que l'on peut obtenir une amélioration sensible des performances de reconnaissance. Nous appliquons cette technique à la comparaison d'une séquence de test codée en MIC différentiel adaptatif (MICDA) avec une séquence de référence normale. Nos études préliminaires montrent qu'on peut améliorer de façon significative la robustesse et les performances des systèmes de reconnaissance de mots isolés, par une extraction de paramètres et une stratégie de comparaison dépendant du signal.

Keywords. isolated word recognition, signal-dependent matching.

1. Introduction

The objective of this paper is to present a new approach to the design of an isolated word speech recognition (IWSR) system. An adaptive matching based on the nature of the input speech segment is proposed here. We call this strategy as signal-dependent matching. To reduce the computational complexity in the implementation of the proposed method, we adopt a two stage approach to matching. In the first stage the warping path between test and reference utterances is determined. In the second stage the distance between the utterances is computed along the path. To measure improvement in recognition performance, a new index of performance is defined. The new index reflects the characteristics of the distance matrix for a given vocabulary rather than the properties of the confusion matrix. The performance of the signal-dependent matching is better than the standard dynamic time warping (DTW) matching algorithm for confusable as well as nonconfusable sets of vocabulary.

The three major components of an automatic speech recognition system are: parameter extraction, creation of reference templates and matching algorithm. Extraction of parameters is required to provide a convenient way of representing and storing the information contained in speech data, which otherwise is too difficult to handle computationally for matching purposes. A variety of parameters are suggested in the literature. Comparative studies of performance of different parameter sets are also reported in [1], [2]. Using a training set of utterances of the words in the vocabulary, reference parameters for each word are created. These parameters reflect important speech-dependent characteristics of the word. In recognition mode, the unknown test utterance, represented by its parameters, is matched with the reference for each word to determine the test word. Matching strategy determines the way the unknown test utterance is compared with the stored references. The two main issues in matching are the nonlinear time registration of the test utterance with a reference word and the distance computation between a test frame and a reference frame. Dynamic time warping (DTW) algorithms have been successfully used for nonlinear time registration [3].

Several studies have been made to optimize separately each of the three components of an IWSR system (Ref: [1], [2] for signal parameter optimization, [4] for reference template creation, [3], [5], [6] for DTW parameter optimization). These studies address the problem mainly from the overall

recognition accuracy point of view, testing the system for large sets of data. Residual errors are attributed to improper end point detection, background noise, etc. Solutions like improved endpoint detection [7], noise suppression or normalization [8] have been suggested. Recently, some studies are reported where different weightages are given to different segments of speech [9], [10]. Although these recent methods appear to be the right approach to isolated word recognition, no major attempt has been made to analyse the limitations of the existing IWSR systems and to determine the causes of the residual error. Moreover, the percentage error as a measure of performance does not adequately describe the sensitivity of the system for variations in the input speech conditions and for variations in the design parameters of the systems. This aspect becomes significant because the present IWSR systems are not still completely error free. In some cases the performance of a system is highly dependent on the signal condition as well as the matching strategy. The purpose of this paper is to examine various factors limiting the performance of an IWSR system. We also examine some possible solutions to the problems and implementation details of the solutions. In particular, we develop a signal-dependent matching algorithm, which takes into account even some distortions in the input speech. We propose a new method of evaluating the system performance at various stages of development of the algorithm.

The paper is organized as follows. In Section 2 some basic limitations of the existing IWSR systems are discussed. The problems due to endpoint detection, fluctuations in the data and matching are discussed in detail. We propose a new method of evaluating the performance of recognition systems in Section 3. The method is developed to reflect the reliability and robustness of recognition rather than percentage accuracy. In Section 4 we explore the possibility of a matching strategy based on the signal knowledge. Some implementation details of the signal-dependent matching using a two-stage approach are also discussed in this section. Finally in Section 5 we indicate the scope of the signal-dependent matching approach in improving the performance of an IWSR system.

2. Some limitations of existing IWSR systems

Presently most IWSR systems claim a recognition accuracy greater than 95% [11]. The recognition accuracy depends on various design choices like size and nature of vocabulary, speaker dependence, background noise, etc. For a given specification, there are other factors that limit the performance of the present systems. These factors relate to the assumptions on the signal prior to DTW operation as well as issues involved in the DTW operation itself. Specifically, we discuss here some limitations imposed by endpoint detection, fluctuations in the data and matching strategy.

2.1. Endpoint detection

Lamel et al. [7] report that recognition reliability and computational complexity of an IWSR system are greatly affected by the accuracy of detection of endpoints of an utterance. The problem of discriminating speech from background noise is not trivial, except in the case of high signal to noise ratio situations. It is extremely difficult to determine the endpoints of an utterance which contains weak fricatives, weak plosive bursts or nasals at the beginning and end of the utterance. Rabiner and Sambur [12] give an algorithm for endpoint detection based on energy and zerocrossings and Lamel et al. [7] suggest some improvements to the algorithm.

Despite these efforts, the endpoints detection still remains a critical issue in IWSR, if high recognition accuracy is desired. One reason for this difficulty is that the endpoints are not always precisely defined. In some cases there may not be any unique endpoints and hence even manual identification of the points from the speech waveform may not be possible. Thus refinements in the endpoint detection becomes a futile exercise in such cases. The only alternative is to accept the

ambiguity in the endpoints and try to design matching algorithms to take care of the ambiguity.

2.2. Fluctuations in the data

Usually speech signal is processed in blocks of uniform segments, called frames, and each frame is represented by some extracted parameters for the purpose of data reduction and comparison. Relative merits of various parametric representations for IWSR have been reported in literature [1], [2]. Generally random fluctuations can be observed in any parameter contour representation of speech data. These fluctuations are a result of natural variations occurring at the speech signal level itself or they may be introduced during various stages of speech signal processing. The natural fluctuations are contributed by speaker variations, background noise, transducer characteristics etc. The fluctuations introduced during signal processing are due to effects of finite block size, frame rate, windowing, averaging, etc. The fluctuations in the data at any level of representation give rise to errors in the pattern comparison process. Regions of wild fluctuations are likely to produce unpredictable distances, which may override the relatively small distances obtained for the remaining matching regions, thus affecting the final decision in the comparison.

One method of reducing the effects of fluctuations is by smoothing the parameter contours. Linear and nonlinear smoothing techniques that smooth out local fluctuations and preserve the relevant changes in parameter contours are described in [13]. Other methods to reduce fluctuations are parameter reduction and parameter averaging. Parameter reduction refers to the process of computing a reduced set of parameters like mel frequency cepstrum, linear prediction coefficients, etc. [1]. Averaging is used to obtain a new set of parameters from the original set by appropriately grouping the parameters and finding the average for each group. This is especially useful to reduce the number of parameters in the spectral domain.

It is to be noted that all the three techniques for reducing the effects of fluctuations (viz smoothing, reduction, averaging) do not consider the reasons for fluctuations. As a consequence it is possible that some of these techniques may average out the relevant information also. An algorithm that makes use of signal knowledge and decides adaptively the compensating strategy for the fluctuations will yield much better results compared to the present techniques. If this compensation strategy can be incorporated in the matching stage, the arbitrariness at the smoothing stage can be avoided.

2.3. Matching strategy

As mentioned earlier, matching strategy refers to the operation of comparison of a test utterance with a stored reference word through DTW algorithm by computing distances between the parameters of the test and reference frames. Several modifications of the DTW algorithm have been proposed to improve the recognition performance. Myers *et al.* [3] give a comparative evaluation of different versions of DTW algorithms. The major limitation of these versions is that the basic algorithm does not permit any change in the matching strategy during comparison. The need for such changes arises due to the fact that different categories of speech segments, namely, silence, fricative, voiced, etc., do not require the same kind of matching. For example, a silence frame when matched with another silence frame yields sometimes arbitrarily large distance due to noise in the data. It is obvious that once we know that a frame belongs to the silence category, its parameters do not contain any useful information and hence it does not require the same procedure for distance computation as for voiced category, for example. But the present DTW algorithms treat all categories alike. The algorithms are not flexible enough to adapt to the characteristics of the input signal.

3. A method for performance evaluation

Performance of speech recognition systems is generally described by a confusion matrix or is

expressed by a single number like percentage recognition error. But it is obvious that a confusion matrix does not show how close the confused words are to the correct ones. The percentage error gives the overall recognition performance only when a large set of test data is available. These measures are not suitable to compare the system performance for different signal-dependent matching schemes during design stage, especially, when only a small test data set is available for experimentation. For example, the percentage error cannot distinguish the distance matrices given in Table 1 and Table 2a. These matrices are obtained without and with the use of a signal-dependent matching respectively, in a recognition experiment. Each matrix is obtained by comparing a test data set, consisting of one sample per word in the vocabulary, with a reference set containing one reference for each word. Both the distance matrices yield the same confusion matrix and hence the same percentage error. What is needed is a measure that reflects the characteristics of the distance matrix, so that one can distinguish the performance as reflected by the above distance matrices. We propose a new measure of performance (PI) that is suitable for evaluating a recognition system under different conditions of operation. First the distance matrix is computed for one set of

Table 1

Distance matrix for a digit vocabulary obtained using conventional approach of signal matching

Test word \ Reference word	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	2145	3712	3043	3178	3244	2652	3261	2406	4390	3572
One	2054	967	2143	2157	1893	1716	3330	1986	3094	1342
Two	1721	2177	1078	1546	2445	2513	2443	1832	2358	2681
Three	2656	3438	2487	1565	4138	2858	3267	2919	3180	3492
Four	2137	3066	3086	2754	2998	2472	3549	2691	3937	3099
Five	2375	2221	3038	2893	2302	1526	3620	2508	4115	1856
Six	4042	6132	4236	4686	5786	4706	1600	2748	3136	6374
Seven	2404	3764	2716	2894	3624	2565	2260	1475	4294	3836
Eight	3711	3948	3087	3362	4330	3984	1534	3005	1297	4540
Nine	2116	1484	2800	2871	1959	1416	3394	2255	3559	1355

Table 2a

Distance matrix for the digit vocabulary obtained using signal-dependent matching

Test word \ Reference word	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	885	2011	1568	2248	1811	1443	1508	1068	2094	2026
One	920	328	1086	1296	660	598	1292	837	1331	642
Two	856	997	354	803	1325	1040	793	646	858	1101
Three	1650	1429	1077	522	2347	1550	1897	1589	1471	1474
Four	1142	1570	1805	2004	1326	1379	1956	1321	2345	1687
Five	1417	979	1892	2189	1178	565	1833	1427	2120	968
Six	1498	2105	7307	2083	1431	1738	475	1149	1812	2542
Seven	1038	1773	1134	1476	1702	1142	1011	486	1758	1827
Eight	920	1066	598	965	1056	1101	589	904	306	1179
Nine	1269	836	1617	1828	1016	635	1517	1196	1829	580

test data. The distances along a row of the matrix correspond to the distances of a test utterance with the reference for each word. The matrix is normalized by dividing all the distances along a row by the corresponding diagonal element value and multiplying the result with 100. This way all the diagonal distances are set to 100 and all the off-diagonal distances are normalized with respect to their diagonal distances. The normalized distance matrix for the data in Table 2a is shown in Table 2b.

From the normalized distance matrix a performance index matrix is derived using a mapping function shown in Fig. 1. The performance index matrix for the data given in Table 2b is shown in Table 2c. The average value of all the off-diagonal elements in the performance index matrix gives the PI of the overall system. The PI for the data in Table 2c is 93.22. The PI for the data in Table 1 is 84.50. These values of PI clearly show that the system corresponding to the matrix in Table 2a is better than the one corresponding to the matrix in Table 1. Note that the PI, as defined here, is suitable mainly to compare the performance of recognition schemes for a given speech data.

The PI is a function of the test data set also. If a large number of test data sets are available, the PI for each set is separately determined and the

Table 2b

Normalized distance matrix for the data in Table 2a

Test word \ Reference word	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	100	227	177	254	204	163	170	120	236	228
One	280	100	331	395	201	182	393	255	405	195
Two	241	281	100	226	374	293	224	182	242	311
Three	316	273	206	100	449	296	363	304	281	282
Four	86	118	136	151	100	103	147	99	176	127
Five	250	173	334	387	208	100	324	252	375	171
Six	315	443	1538	438	301	365	100	241	381	535
Seven	213	364	233	303	350	234	208	100	361	375
Eight	300	348	195	315	345	359	192	295	100	385
Nine	218	144	278	315	175	109	261	206	315	100

Table 2c

Performance index matrix corresponding to the normalized distance matrix given in Table 2b

Test word \ Reference word	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	10	100	100	100	100	98	100	40	100	100
One	100	10	100	100	100	100	100	100	100	100
Two	100	100	10	100	100	100	100	100	100	100
Three	100	100	100	10	100	100	100	100	100	100
Four	3	36	71	86	10	13	82	9	100	54
Five	100	100	100	100	100	10	100	100	100	100
Six	100	100	100	100	100	100	10	100	100	100
Seven	100	100	100	100	100	100	100	10	100	100
Eight	100	100	100	100	100	100	100	100	10	100
Nine	100	79	100	100	100	19	100	100	100	10

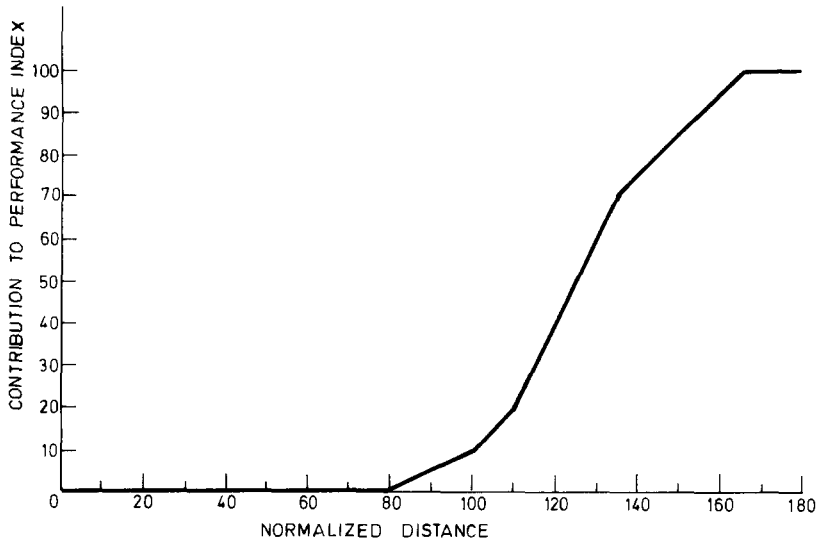

Fig. 1. Mapping function from normalised distance matrix to the performance index matrix.

average of these PI values is taken as the performance index of the system. This method of performance evaluation has the advantage that it reflects the 'goodness' of the distance matrix, thus giving an indication of the reliability of the recognition scheme. Recognition error represents the number of mismatches, whereas the PI represents the severity of the mismatches. The PI is also useful in estimating the confusability of a given vocabulary. If the PI is low, it is an indication that the words in the vocabulary are confusable. But very low values of PI (<20) have no meaning because the various thresholds chosen for the mapping function are reasonable only when the recognition system is performing well.

The approach used to derive the PI is purely heuristic. A detailed analysis of a large number distance matrices for different sets of vocabulary will help in arriving at better mapping functions and hence a more logical performance index measure. But still the method suggested here serves the purpose of comparing several signal-dependent matching schemes. One major advantage of this measure is that the performance can be evaluated using a small set of data, and is thus ideally suited for evolving matching strategies using modest computational facilities.

4. Signal-dependent matching

4.1. Proposed scheme

The basic limitations of the present DTW algorithms can be overcome by making use of the signal knowledge in parametric extraction and matching strategy. In this paper we consider the use of signal knowledge in determining an adaptive matching strategy.

The process of making the matching algorithm adaptive based on the signal knowledge consists of two distinct steps. The first step is to derive the signal knowledge and the second step is to use the knowledge in the matching algorithm. Signal knowledge can be derived in a variety of ways. For example, from amplitudes and zerocrossings of a signal waveform it is possible to identify silence and transition regions of an utterance. Alternatively, spectral parameters may serve as a convenient means of representing the signal characteristics from which signal knowledge can be extracted. The main difficulty in developing the matching strategy is to know what knowledge of the signal is required. To develop an adaptive matching algorithm, each utterance is considered as a sequence of segments or frames. Based on the

signal characteristics of each frame, the parameters to represent the frame and a dissimilarity measure between frames have to be evolved.

From implementation point of view, changes in the parameter set and in the matching strategy are ultimately intended to give due weightage to the distance computed for each test frame. This can be implemented by creating a weight function for each utterance. The weight function, consisting of tuples for each frame, will indicate the parameter set to be used for each frame and the weighting factor for the computed distance for that frame. The values for each tuple should be determined using signal knowledge.

It is to be noted that the weight function should depend only on the test utterance and not on the reference data. If the weight function depends on reference data also, then the distances between a given test utterance and different reference words do not have any common basis for comparison. For example, some form of normalization of the distance will have to be made to compare the distances of a test utterance with different reference words, if different sets of test frames are used while matching with different reference words, as in the unconstrained endpoint DTW algorithm [6]. Normalization of distances would also be required if the number of times a test frame is used in the distance computation is dictated by the reference word being matched. For this reason, DTW algorithms which use symmetric comparison as in [5] or any DTW algorithms which skips test frames based on reference data are not suitable for signal-dependent matching.

Assuming that a weight function is available, the signal-dependent matching can be implemented either by direct method or by a two stage approach. In the direct method the signal-dependent matching is implemented along with the conventional DTW algorithm. In this method the operations of fixing a warping path and computation of the distance are performed simultaneously according to the weight function. But this approach may involve significantly large computational effort because each test frame has its own parameter set and its associated procedure for distance computation. The warping path is guided by the distance computed and the distance is computed along the warping path. Thus the two operations appear to be interdependent. To reduce the computational effort, we examine the feasibility of another method, called a two stage approach.

In the two stage approach the two basic operations of the DTW algorithm are treated as independent. This approach is based on the assumption that, if different parameter sets are available to represent speech data, there should not be any significant differences among the warping paths obtained by the different parameter sets. In the first stage the warping path is determined using computationally simple parameter set in the normal DTW algorithm. At this stage a simple distance computation is used to determine the warping path. In the second stage the interframe distances along the path are computed according to the weight function. In other words, different parameter sets, distance measures and weights can be used for each test frame based on the knowledge of the signal. The distance computation for each test frame could be very complex without increasing the overall complexity in the matching. It is to be noted that computational complexity of the two stage approach is almost same as for the normal DTW algorithm.

Although the two stage approach is definitely not optimal, the loss of performance can be more than compensated by incorporating fairly sophisticated decision rules derived from the signal knowledge.

4.2. Description of parameters

In this section we shall illustrate the performance improvement due to signal-dependent matching in IWSR system. For this purpose we consider a basic IWSR system. The system consists of parameter extraction and matching stages. Speech data consists of two repetitions of a small subset of alphadigit vocabulary. One of the repetitions is used for reference data and the other for test data.

The speech signal is sampled at 10 KHz and digitized and stored as 16 bit numbers. Each utterance is divided into nonoverlapping frames of size 256 samples. Each frame is multiplied by a Hamming window function and its discrete Fourier transform is computed. The spectrum is reduced to 16 log spectral values on an approximate melfrequency scale as given in [1]. Each test utterance is compared with each of the reference words using DTW algorithm.

The DTW algorithm is similar to the asymmetric algorithm proposed by Itakura [14]. The distance between a test frame and a reference frame is computed by summing the absolute difference between the corresponding log melspectral values.

The DTW algorithm given in [14] forces the begin and end frames of a test utterance to match with the begin and end frames of each reference word. But it is well known that begin and end of an utterance cannot always be determined uniquely. A better approach, therefore, is to allow some flexibility at the beginning and ending of each utterance during matching. In particular, the first test frame should be allowed to match any one of the first few frames of a reference word. Similarly, the last test frame should be allowed to match with any one of the last few frames of the reference word. This scheme will not permit any test frame to be skipped and also each test frame is used only once in the final minimum distance computation. The search region which allows this end point flexibility is shown in Fig. 2. The region is defined by two parallel lines with slope 1/2 in the (m, n) plane, one passing through (1, 1) and the other through (M, N). To prevent skipping of too many reference frames, the candidate reference frames for matching the first and the last test frames are restricted to a 50 msec window.

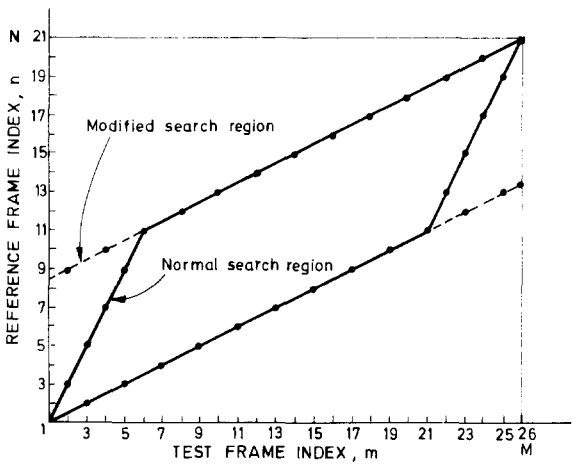

Fig. 2. Modified search region to allow endpoint flexibility in dynamic time warping.

4.3. *Approximate warping path using standard DTW algorithm*

The objective of our two stage approach is to reduce computational effort involved in implementing the signal-dependent matching. The first stage consists of determining the warping path. A good approximation to the warping path can be obtained using a few parameters and a simple distance computation. We demonstrate this by considering warping paths obtained for several choices of parameter sets. Five parameter sets consisting of 16, 8, 4, 2 and 1 parameters are denoted as P16, P8, P4, P2 and P1. The parameter set P16 correspond to the 16 log melspectral values. The other sets P8, P4, P2 and P1 are derived from P16 by averaging the adjacent 2, 4, 8 and 16 log melspectral values for each frame respectively.

Fifty word matching experiments were carried out to decide suitable parameter set for warping path determination. Two sets of vocabulary are considered for this investigation: One confusable set consisting of the words, A, B, C, D, E and one nonconfusable set consisting of the words *Zero, One, Two, Three, Four.* Warping paths were obtained for each of the 50 comparisons using the four parameter sets P16, P8, P4 and P2. In every comparison we found that there are no significant differences among the warping paths obtained by the different parameter sets. Fig. 3 shows typical warping paths obtained in these comparisons. The experiment has shown that warping path can be determined using two parameters per frame.

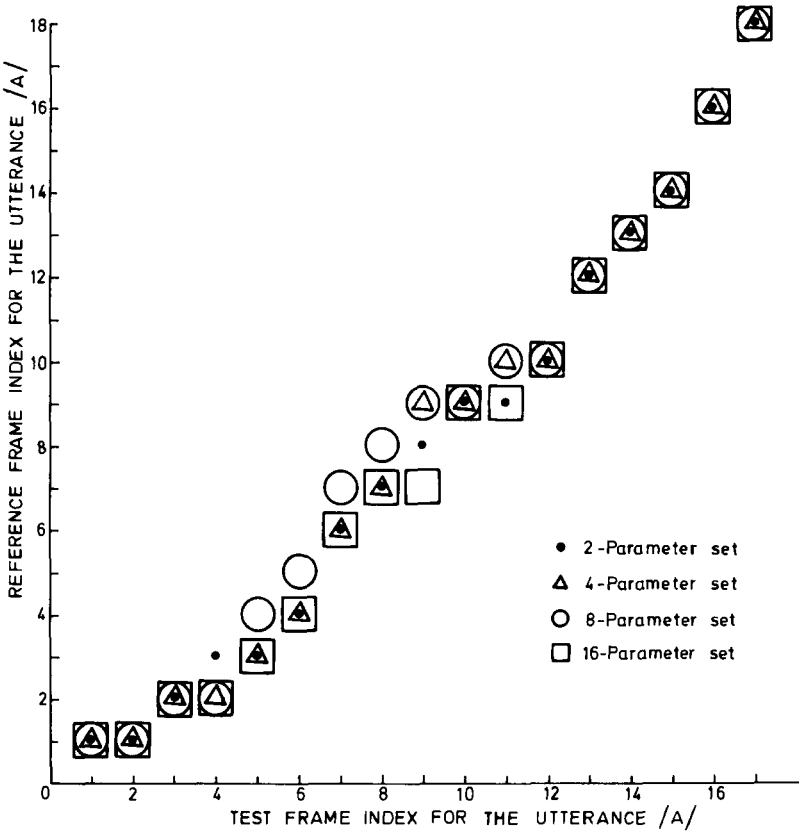

Fig. 3a. Warping paths for different parameter sets (Example 1).

4.4. *Distance computation using weight function*

The signal-dependent matching distance between a test utterance and a reference word is computed along the predetermined warping path using a weight function. The weight function consists of a weight vector or a tuple for each test frame. The first component of the vector indicates the parameter set to be used for that frame. Usually the parameter set automatically decides the type of distance computation to be adopted. The second component of the vector represents a scale factor for the computed distance. To derive the weight vectors, the frames of the test utterance are classified into voiced/unvoiced/silence (V/U/S) classes. The silence and nonsilence classification can be done using relative energy levels of different frames. The high and low frequency components of the spectrum can be used in distinguishing voiced and unvoiced frames. It should be noted that the classification of speech frames into different categories need not be very accurate. A simple algorithm for this classification was used in our experiments.

The algorithm consists of two steps. In the first step, the peak energy frame of an utterance is identified. The frames in the end portions of the utterance with energy level (in dB) less than 80% of the peak energy level are treated as candidates for silence region. In the second step, the signal is scanned from the end frames towards inside. The frame which shows a 30% increase of energy level as compared to the average energy level of all the previous frames is taken as the boundary frame of the nonsilence portion. The actual nonsilence portion is selected from these two steps, whichever gives maximum coverage.

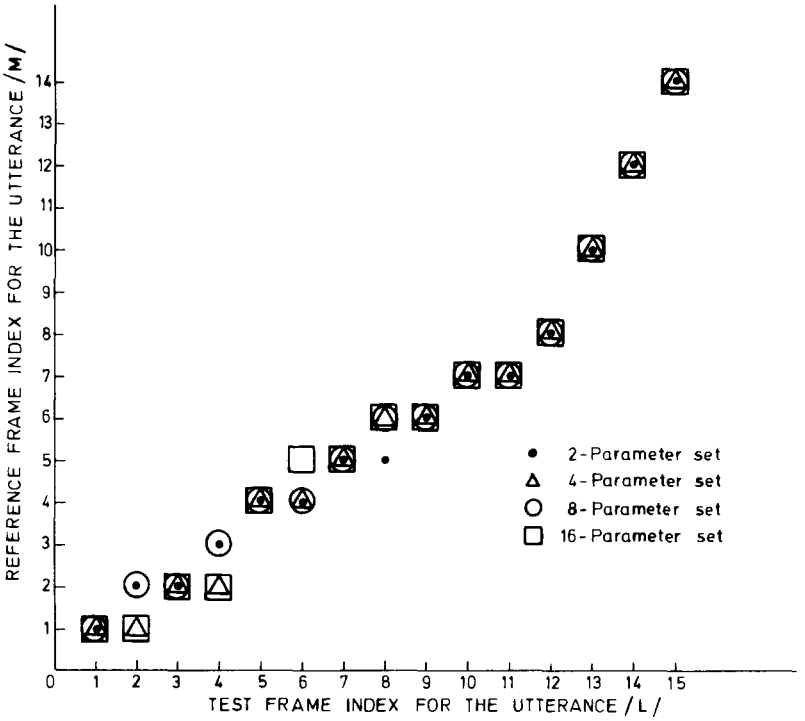

Fig. 3b. Warping paths for different parameter sets (Example 2).

The nonsilence frames are classified into voiced and unvoiced categories using the relative levels of the low frequency and high frequency components. A higher value of the low frequency component indicates that it is a voiced frame.

The parameters used to represent a frame are varied depending on the class of the test frame. For all nonvoiced frames all the 16 log melspectral values are used as parameters. For voiced frames, the deviation of the log melspectral values from the mean value for each frame are considered as parameters. These zero-mean parameters provide some kind of energy normalization. Energy normalization is generally required for voiced frames due to significant fluctuations in energy from repetition to repetition. It is important that, while matching, the same parameters have to be used for the test and reference frames.

The scale factor for each frame is determined based on the energy level and the class of the frame. Generally high energy frames are given higher weightage than low energy frames. Weightage is reduced for transition frames between any two of the three classes, because it is difficult to reliably reproduce the characteristics of the transition frames for each repetition. In our experiments the following weights are given for different classes of frames.

Let x be the maximum energy level (in dB) of a frame in the test utterance. Then for energy level less than $0.8x$, the weight is 0.2; for energy level in the range $0.8x$ to x, the weight is 1.0; for all transition frames, the weight is 0.5. Note that the weightages are somewhat arbitrary. There is scope for refining these values based on large number of experiments.

4.5. Experiments using signal-dependent matching

Experiments were conducted to study the performance of the signal-dependent matching. Table 3 and Table 4 show the improvement in PI using the weight vector matching over the conventional DTW algorithm. Although the two-stage approach

Table 3
Improvement of performance index using signal-dependent matching over the conventional DTW algorithm

Vocabulary: /F, H, L, M, N, X/	
Performance Indices obtained using:	
a) conventional approach	56.34
b) two-stage approach	62.38
c) addition of weight-vector	78.14
d) addition of zero-mean	79.21

alone does not guarantee any improvement, the use of signal-dependent matching has significantly improved the PI for different types of vocabulary.

5. Summary and conclusions

5.1. Summary

In this paper we have proposed a new method to design IWSR systems. The method uses signal knowledge of the test utterance to adaptively change the matching strategy. We have shown that the signal-dependent matching improves recognition performance significantly.

The performance index (PI) measure proposed in this paper is suitable when a small amount of test data is available for recognition system development. The PI gives an indication of the reliability and robustness of the recognition system. It enables comparison of several design choices to arrive at a good design. We have demonstrated the need for signal-dependent matching for improving the performance. The warping path in a DTW algorithm does not depend critically on the parameter set used. This property was exploited to reduce computational effort in the implementation of the signal-dependent matching by proposing a two-stage approach for distance computation between a test utterance and a reference word. Confusable and nonconfusable sets of vocabulary were studied separately to demonstrate that the value of PI gives also an indication of the confusability of the vocabulary. Finally, we show in [15] that signal-dependent matching can be effectively used to counter the effects of distortions in the input speech, which cannot otherwise be handled easily due to unpredictable nature of distortions.

5.2. Limitations

Our choice of log melspectral parameters is somewhat arbitrary. We have not attempted to optimize the parameter set to represent each frame. However, the log melspectral values have the advantage that spectral information can be represented compactly. Moreover, the variance in spectral data is significantly reduced due to averaging process used in deriving the melspectral parameters. The derivation of the weight vector, warping path and distance computation also have not been optimized.

Finally, the mapping function used in deriving the PI is purely empirical and it needs to be refined based on large number of recognition experiments on a variety of speech data.

5.3. Further work

What is reported in this paper is only an exploratory study. It is possible to improve the perform-

Table 4
Improvement of performance index using signal-dependent matching for different sets of vocabulary

Vocabulary	/B, C, D, E, G, P, T, V, Z/	/A, J, K/	/0, 1, ... 9/
Conventional approach	70.25	85.63	84.50
Two stage approach with weight vector and zero-mean	84.26	97.33	93.47

ance of a recognition system by incorporating more sophistication in the matching algorithm through the weight vector concept. Without increasing the computational complexity, it is possible to use different parameter sets and distance computations for different test frames based on signal knowledge. For this, it is necessary to use relevant additional knowledge from the signal. It is also necessary to determine what is the relevant knowledge and how to extract it from the data.

In conclusion, we can say that no significant improvement in the performance of the current IWSR systems can be achieved by individually attacking the problems of parameter optimization, DTW algorithm and endpoint detection. Signal-dependent matching along the lines indicated in this paper offers much scope for improving the reliability and robustness, besides accuracy, of a recognition system. Moreover, it appears that signal-dependent matching is a natural way of processing speech, especially for machine recognition.

Acknowledgement

The authors wish to thank the reviewers for many valuable suggestions to improve the presentation of the paper.

References

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-28, Aug. 1980, pp. 357–366.

[2] G.M. White and R.B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-24, Aug. 1976, pp. 289–295.

[3] C. Myers, L.R. Rabiner and A.E. Rosenberg, "Performance tradeoffs in DTW algorithms for isolated word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-28, Dec. 1980, pp. 622–635.

[4] L.R. Rabiner, "On creating reference templates for speaker independent recognition of isolated words", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, Feb. 1978, pp. 34–42.

[5] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, Feb. 1978, pp. 43–49.

[6] L.R. Rabiner, A.E. Rosenberg and S.E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, Dec. 1978, pp. 575–582.

[7] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, "An improved endpoint detector for isolated word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-29, Aug. 1981, pp. 777–785.

[8] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, Vol. 67, Dec. 1979, pp. 1587–1604.

[9] K. Elenius and M. Blomberg, "Effect of emphasizing transitional or stationary parts of the speech signal in discrete utterance recognition system", *Proceedings of ICASSP-82 Conference*, Paris, France, May 1982, pp. 535–538.

[10] L.J. Siegel and A.C. Bessey, "Voiced/unvoiced/mixed excitation classification of speech", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-30, Jun. 1982, pp. 451–460.

[11] 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.

[12] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell Syst. Tech. J.*, Vol. 54, No. 2, February 1975, pp. 297–315.

[13] L.R. Rabiner, M.R. Sambur and C.E. Schmidt, "Applications of a non-linear smoothing algorithm to speech processing", *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, No. 6, December 1975, pp. 552–557.

[14] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-23, Feb. 1975, pp. 67–72.

[15] B. Yegnanarayana, Sarat Chandran and Anant Agarwal, "On improvement of performance of isolated word recognition for degraded speech", *Signal Processing*, Vol. 7, No. 2, 1984, pp. 175–183.