COMPARISON OF PARAMETRIC REPRESENTATIONS FOR MONOSYLLABIC WORD RECOGNITION IN CONTINUOUSLY SPOKEN SENTENCES*

Steven B. Davis+ and Paul Mermelstein++

Abstract. Several parametric representations of the acoustic signal were compared as to word recognition performance in a syllable-oriented continuous speech recognition system. The vocabulary included many phonetically similar monosyllabic words, therefore the emphasis was on ability to retain phonetically significant acoustic information in the face of syntactic and duration variations. For each parameter set (based on a mel-frequency cepstrum, a linear frequency cepstrum, a linear prediction cepstrum, a linear prediction spectrum, or a set of reflection coefficients), word templates were generated using an efficient dynamic warping method, and test data were time registered with the templates. A set of ten mel-frequency cepstrum coefficients computed every 6.4 ms resulted in the best performance, namely 96.5% and 95.0% recognition with each of two speakers. The superior performance of the mel-frequency cepstrum coefficients may be attributed to the fact that they better represent the perceptually relevant aspects of the short-term speech spectrum.

1. INTRODUCTION

The selection of the best parametric representation of acoustic data is an important task in the design of any speech recognition system. The usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences. When a significant amount of reference information is stored, such as different speakers' productions of the vocabulary, compact storage of the information becomes an important practical consideration.

---

*To appear in IEEE Transactions on Acoustics, Speech and Signal Processing.

+Now at Signal Technology, Inc., 15 W. De La Guerra St., Santa Barbara, CA 93101

++Now at Bell-Northern Research and INRS-Telecommunications, University of Quebec, 3, Place du Commerce, Nuns' Island, Verdun, Quebec, Canada H3E 1H6

Acknowledgement. This material is based upon work supported by NSF Grant BNS 7682023 to Haskins Laboratories. Drs. Frank Cooper and Patrick Nye participated in numerous discussions of the experimental program, and their contribution is greatly appreciated.

[HASKINS LABORATORIES: Status Report on Speech Research SR-61 (1980)]

The choice of a basic phonetic segment bears closely on the representation problem because the decision to identify an unknown segment with a reference category is based on the parameters within the entire segment. The number of different reference segments is generally smaller than the number of possible unknown segments, and therefore the step of identifying an unknown with a reference entails a significant loss of information. One can minimize the loss of useful information by examining different parametric representations in the framework of the specific recognition system under consideration. However, since the choice of a segment is so basic to the decision as to what acoustic information is useful, the result of such a comparative examination of different representations is directly applicable only to the specific recognition system, and generalization to differently organized systems may not be warranted.

Fujimura (1975) and Mermelstein (1975b) discussed in detail the rationale for use of syllable-sized segments in the recognition of continuous speech. The goal of the experiments reported here was to select an acoustic representation most appropriate for the recognition of such segments. The methods used to evaluate the representations were open testing, where the training data and test data were independently derived, and closed testing, where these data sets were identical. In each case, the same speaker produced both the reference and test data, which included the same words in a variety of different syntactic contexts. Although variation between speakers is an important problem in its own right, attention is focused here on speaker dependent representations to restrict the different sources of variation in the acoustic data.

White and Neely (1976) showed that the choice of parametric representations significantly affects the recognition results in an isolated word recognition system. Two of the best representations they explored were a 20-channel bandpass filtering approach using a Chebychev norm on the logarithm of the filter energies as a similarity measure, and a linear prediction coding approach using a linear prediction residual (Itakura, 1975) as a similarity measure. From the similarity of the corresponding results, they concluded that bandpass filtering and linear prediction were essentially equivalent when used with a dynamic programming time alignment method. However, that result may be due to the absence of phonetically similar words in the test vocabulary.

Because of the known variation of the ear's critical bandwidths with frequency (Feldtkeller & Zwicker, 1956; Schroeder, 1977), filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Pols (1977) showed that the first six eigenvectors of the covariance matrix for Dutch vowels of three speakers, expressed in terms of 17 such filter energies, accounted for 91.8% of the total variance. The direction cosines of his eigenvectors were very similar to a cosine series expansion on the filter energies. Additional eigenvectors showed an increasing number of oscillations of their direction cosines with respect to their original energies. This result suggested that a compact representation would be provided by a set of mel-frequency cepstrum coefficients. These cepstrum coefficients are the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale.[1]

A preliminary experiment (Mermelstein, 1976) showed that the cepstrum coefficients were useful for representing consonantal information as well. Four speakers produced 12 phonetically similar words, namely "stick," "sick," "skit," "spit," "sit," "slit," "strip," "scrip," "skip," "skid," "spick," and "slid." A representation using only two cepstrum coefficients resulted in 96% correct recognition of this vocabulary. Given these encouraging results, it became important to verify the power of the mel-frequency cepstrum representation by comparing it to a number of other commonly used representations in a recognition framework where the other variables, including vocabulary, are kept constant.

This paper compares the performance of different acoustic representations in a continuous speech recognition system based on syllabic units. The next section describes the organization of the recognition system, the selection of the speech data, and the different parametric representations. The following section describes the method for generating the acoustic templates for each word by use of a dynamic warping time alignment procedure. Finally, the results obtained with the various representations are listed and discussed from the point of view of completeness in representing the necessary acoustic information.

2. The Experimental Framework

A rather simple speech recognition framework served as the testbed to evaluate the various acoustic representations. Lexical information was utilized in the form of a list of possible words and their corresponding acoustic templates, and these words were assumed to occur with equal likelihood. No syntactic or semantic information was utilized. If such information had been present, it could have been used to restrict the number of admissible lexical hypotheses or assign unequal probabilites to them. Thus, in practice, instead of matching hypotheses to the entire vocabulary, the number of lexical hypotheses that one evaluates may be reduced to a much smaller number. This reduction would cause many of the hypotheses phonetically similar to the target word to be eliminated from consideration. Thus the high phonetic confusability of the test data may have resulted in a test environment that is more rigorous than would be encountered in practice.

2.1 Selection of Corpus

The performance of continuous speech recognition systems is determined by a number of distinct sources of acoustic variability, including speaker characteristics, speaking rate, syntax, communication environment and recording and/or transmission conditions. The focus of the current experiments is acoustic recognition in the face of variability induced in words of the same speaker by variation of the surrounding words and by syntactic position. The use of a separate reference template for each different syntactic environment which a word might occupy would require exorbitant amounts of storage and training data. Thus an important practical requirement is to generate reference templates without regard to the syntactic position of the word. To avoid the problem of automatically segmenting complex consonantal clusters, the corpus was composed of monosyllabic target words that were semantically acceptable in a number of different positions in a given syntactic context. Since acoustic variation due to different speakers is a distinctly separate

problem (Rabiner, 1978), it was considered advisable to restrict the scope of these initial experiments by using only speaker dependent templates. That is, both reference and test data were produced by the same speaker.

The sentences were read clearly in a quiet environment and recorded using a high quality microphone. These recording conditions were selected to establish the best performance level that one could expect the recognition system to attain. Environments with higher ambient noise, which may be encountered in a practical speech input situation, would undoubtedly detract from the clarity of the acoustic information and therefore result in lower performance.

The speech data comprised 52 different CVC words from two male speakers (DZ and LL), and a total of 169 tokens were collected from 57 distinct sentences (Appendix A). The sentences were read twice by each speaker in recording sessions separated in time by two months (denoted as DZ1, DZ2, LL1 and LL2). Thus the data consisted of a total of 676 syllables. To achieve the required variability, the selected words could be used as both nouns and verbs. For example, "Keep the hope at the bar" and "Bar the keep for the yell" are two sentences that allow syntactic variation but preserve the same overall intonation pattern. All the words examined carried some stress; the unstressed function words were not analyzed. The target words, all CVC's, included 12 distinct vowels, /i, I, e, ɛ, æ, ɔ, ʌ, U, u, ɝ, a, o/, some of which are normally diphthongized in English. Each vowel was represented in at least four different words, and these words manifested differences in both the prevocalic and postvocalic consonants. The consonants comprised simple consonants as well as affricates but no consonantal clusters.

2.2 Segmentation

An automatic segmentation process (Mermelstein, 1975a) was initially considered as one way of delimiting syllable-sized units in continuously spoken text, but any such algorithm performs the segmentation task with a finite probability of error. In particular, weak unstressed function words sometimes appear appended to the adjacent words carrying stronger stress. Additionally, in this study, a boundary point located for an intervocalic consonant with high sonority may not consistently join that consonant to the word of interest. In order to avoid possible interaction between segmentation errors and poor parametric representations, manual segmentation and auditory evaluation was used to accurately delimit the signal corresponding to the target words. The segmentation, as well as the subsequent analysis and recognition, was performed on a PDP-11/45 minicomputer with the Interactive Laboratory System (Pfeifer, 1977).

In systems employing automatic segmentation, the actual recognition rates can be expected to be lower due to the generation of templates from imperfectly delimited words (Mermelstein, 1978). However, there is no reason to believe that segmentation errors would not detract equally from the recognition rates obtained for the various parametric representations.

2.3 Parametric Representations

The parametric representations evaluated in this study may be divided into two groups, those based on the Fourier spectrum and those based on the linear prediction spectrum. The first group comprises the mel-frequency cepstrum coefficients (MFCC) and the linear-frequency cepstrum coefficients (LFCC). The second group includes the linear prediction coefficients (LPC), the reflection coefficients (RC), and the cepstrum coefficients derived from the linear prediction coefficients (LPCC). A Euclidean distance metric was used for all cepstrum parameters, since cepstrum coefficients are derived from an orthogonal basis. This metric was also used for the RC, in view of the lack of an inherent associated distance metric. The LPC were evaluated using the minimum prediction residual distance metric (Itakura, 1975).

Each acoustic signal was lowpass filtered at 5 kHz and sampled at 10 kHz. Fourier spectra or linear prediction spectra were computed for sequential frames 64 points (6.4 ms) or 128 points (12.8 ms) apart. In each case, a 256 point Hamming window was used to select the data points to be analyzed. (A window size of 128 points produced degraded results).

For the MFCC computations, 20 triangular bandpass filters were simulated as shown in Figure 1. The MFCC were computed as

$$MFCC_i = \sum_{k=1}^{20} X_k \cos[i(k - \frac{1}{2})\frac{\pi}{20}] \quad , \quad i = 1,2,\ldots,M \quad , \tag{1}$$

where M is the number of cepstrum coefficients, and X_k, $k = 1,2,\ldots,20$, represents the log-energy output of the kth filter.

The LFCC were computed from the log-magnitude Discrete Fourier Transform (DFT) directly as

$$LFCC_i = \sum_{k=0}^{K-1} Y_k \cos(\frac{\pi ik}{K}) \quad , \quad i = 1,2,\ldots,M \quad , \tag{2}$$

where K is the number of DFT magnitude coefficients Y_k.

The LPC were obtained from a 10th order all-pole approximation to the spectrum of the windowed waveform. The autocorrelation method for evaluation of the linear prediction coefficients was used (Markel & Gray, 1976). The RC were obtained by a transformation of the LPC which is equivalent to matching the inverse of the LPC spectrum with a transfer function spectrum that corresponds to an acoustic tube consisting of ten sections of variable cross-sectional area (Wakita, 1973). The reflection coefficients determine the fraction of energy in a travelling wave that is reflected at each section boundary.

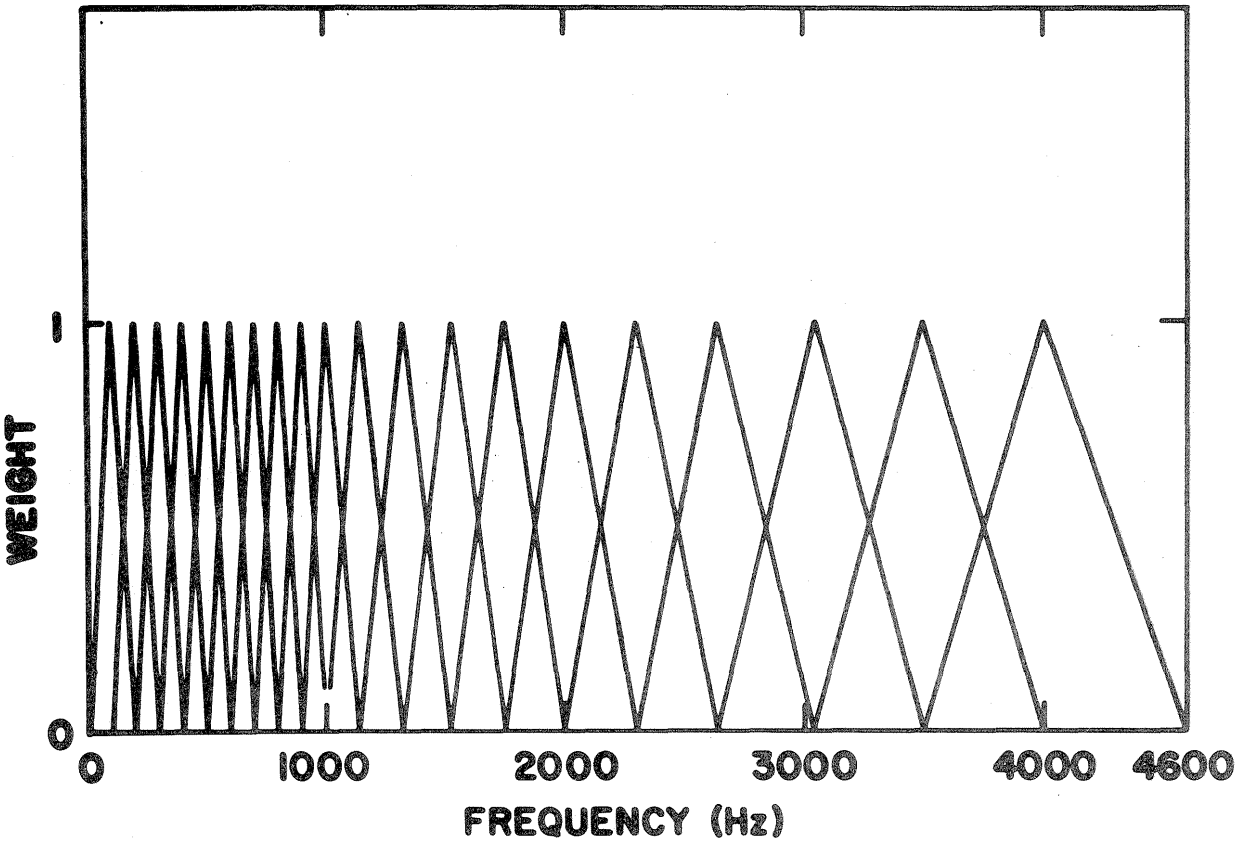


Figure 1: Filters for generating mel-frequency cepstrum coefficients.

The LPCC were obtained from the LPC directly as

$$LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k-i}{i} LPCC_{i-k} LPC_k \quad , \; i=1,2,\ldots,10 \quad . \tag{3}$$

The Itakura metric represents the distance between two spectral frames with optimal (reference) LPC and test $\widehat{LPC}$ as

$$D\,[LPC,\widehat{LPC}] = \log \left| \frac{LPC\ \hat{R}\ LPC^T}{\widehat{LPC}\ \hat{R}\ \widehat{LPC}^T} \right| \quad , \tag{4}$$

where R is the autocorrelation matrix (obtained from the test sample) corresponding to the $\widehat{LPC}$. The metric measures the residual error when the test sample is filtered by the optimal LPC. Because of its asymmetry, the Itakura metric requires specific identification of the reference coefficients (LPC) and the test coefficients ($\widehat{LPC}$). For computational efficiency, the denominator of (4) will be unity if $\hat{R}$ is expressed in unnormalized form. Then if $\hat{r}(n)$ denotes the unnormalized diagonal elements of $\hat{R}$, $r_{LP}(n)$ denotes the unnormalized autocorrelation coefficients from the LPC polynomial, and the logarithm is eliminated, the distance may be expressed as (Gray & Markel, 1976)

$$D[\hat{r},r_{LP}] = \hat{r}(0)r_{LP}(0) + 2 \sum_{i=1}^{10} \hat{r}(i)r_{LP}(i) \quad . \tag{5}$$

3. Generation of Acoustic Templates

The use of templates to represent the acoustic information in reference words allows a significant computation reduction compared to use of the reference tokens themselves. The design of a template generation process is governed by the goal of finding the point in acoustic space that simultaneously minimizes the "distance" to all given reference items. Where the appropriate distance is a linear function of the acoustic variables, this goal can be realized by the use of classic pattern recognition techniques. However, phonetic features are not uniformly distributed across the acoustic data, and therefore perceptually motivated distance measures are nonlinear functions of those data. To avoid the computationally exorbitant procedure of simultaneously minimizing the set of nonlinear distances, templates are incrementally generated by introducing additional acoustic information from each reference word to the partial template formed from the previously used reference words. Given a distance between two tokens, or between a token and a template, the new template can be located along the line whose extent measures that distance. Since only acoustically similar tokens are to be combined into individual templates, one may expect that this procedure will exploit whatever local linearization the space permits.

3.1 Template Generation Algorithms

In one algorithm (Rabiner, 1978), an initial template is chosen as the token whose duration is the closest to the average duration of all tokens representing the same word (Figure 2). Then all remaining tokens are warped to the initial template. The warping is achieved by first using dynamic programming to provide a mapping (or time registration) between any test token and the reference template. Following the notation in Rabiner, Rosenberg, and Levenson (1978), let $T_i(m)$, $0 \leq m \leq M_i$, be a test contour for word replication i with duration M_i, $i=1,2,\ldots,I$, and let $R_1(m) = T_j(m)$ be the initial reference contour, where the duration of the jth token is closest to the average duration. For example, these contours may be vectors of cepstrum coefficients obtained at 10 ms intervals during the word. Then dynamic programming may be used to find mappings $m_i = w_i(n)$, $i=1,2,\ldots,I$, subject to boundary conditions at the endpoints, such that the total distance $D_T(i)$ between test token i and the reference contour is minimal. A distance function D is defined for each pair of points (m,n). Then

$$D_T(i) = \min_{\{w_i(n)\}} \sum_{n=1}^{N} D[R_1(n), T_i(w_i(n))] \quad . \tag{6}$$

With the aid of these mappings, a new reference contour may be defined as

$$R_2(n) = \frac{1}{I} \sum_{i=1}^{I} T[w_i(n)] \quad , \tag{7}$$

and the process is repeated until the distance between the current and previous templates is below some threshold. This procedure is not dependent on the order in which tokens are considered. However, it is computationally expensive to iterate to the final reference contour. Furthermore, there may be cases where there is no convergence (Rabiner, 1978).

A different algorithm can be used for phonetically similar words; this algorithm requires less computation effort and has no convergence problems. Furthermore, the algorithm allows a reference template to be easily updated with an accepted token during verification to allow for word variation over time. In this procedure (Davis, 1979), each successive token is warped with the current template to produce a new template for the next token (Figure 3). For example,

$$\begin{aligned}
R_1(n) &= T_1(n) \quad , \\
R_2(n) &= \frac{1}{2}\,[\,R_1(n) + T_2(w_2(n))] \quad , \\
R_3(n) &= \frac{1}{3}\,[2R_2(n) + T_3(w_3(n))] \quad , \\
&\;\;\vdots \qquad\qquad \vdots \\
R_I(n) &= \frac{1}{I}\,[(I-1)R(I-1,n) + T_I(w_I(n))] \quad .
\end{aligned} \tag{8}$$

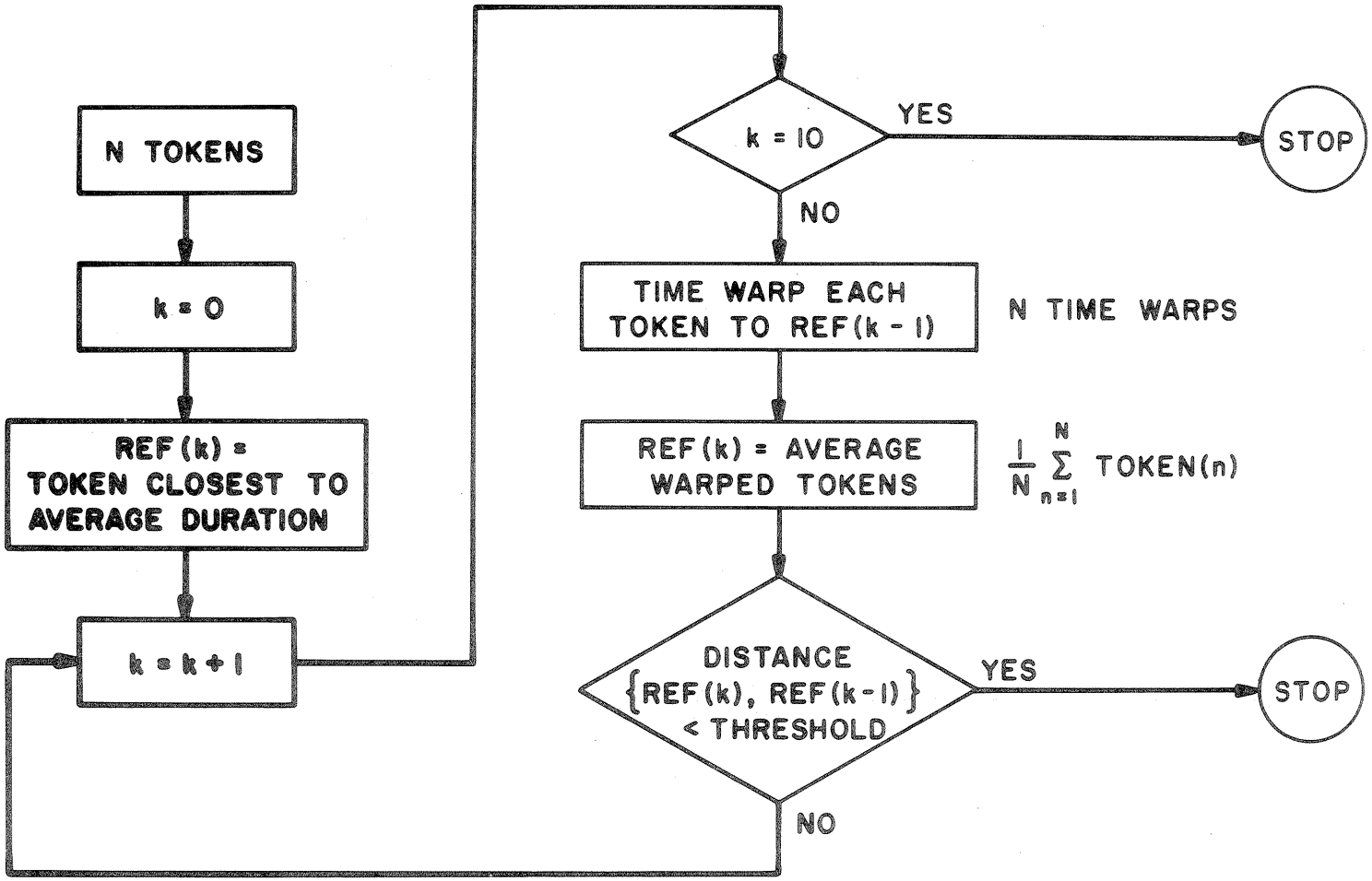


Figure 2: Iterative algorithm for template generation.

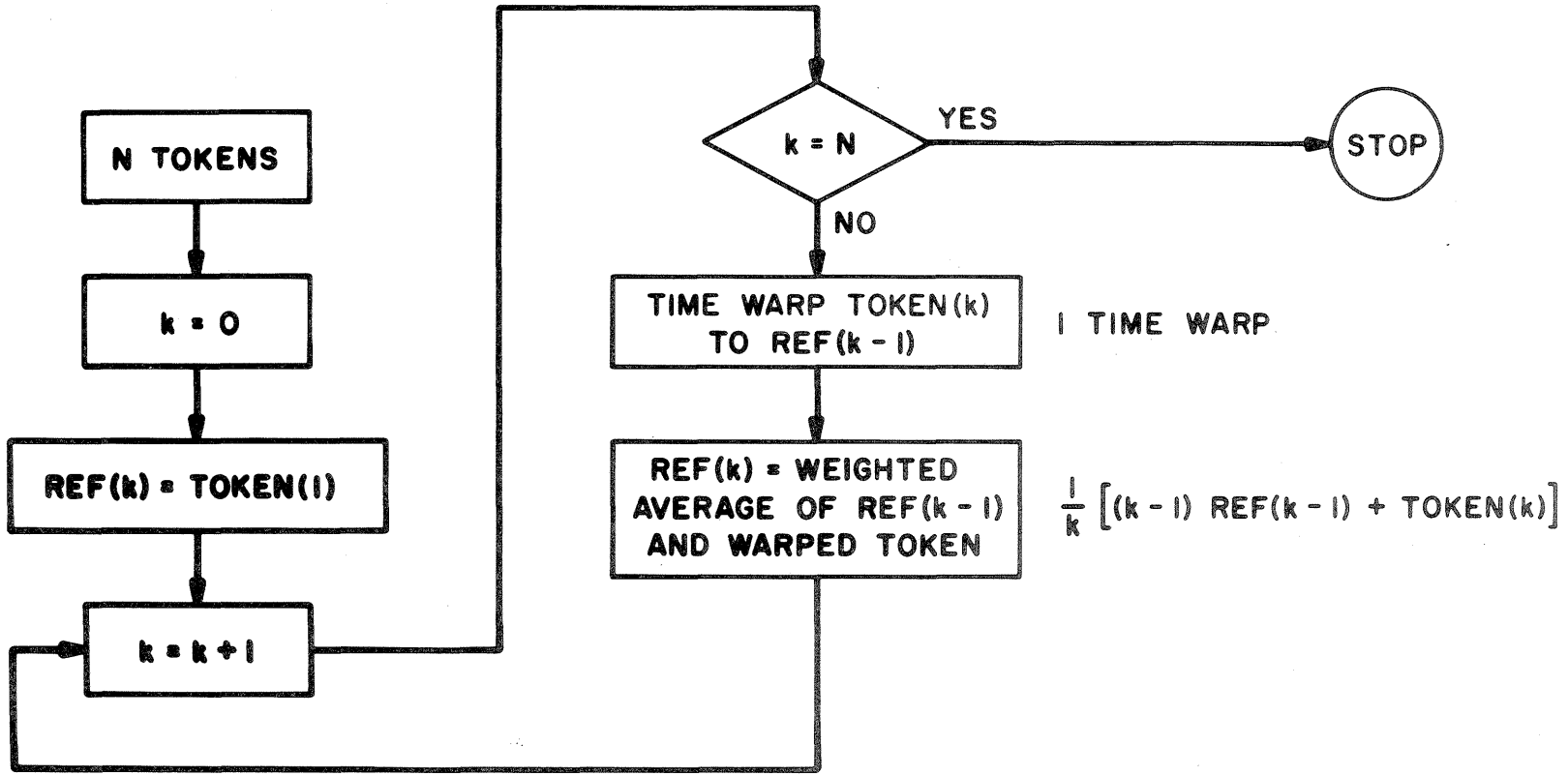


Figure 3: Noniterative algorithm for template generation.

Thus, the process ends with the Ith template.

While this algorithm has computational advantages over the first algorithm, the results become order dependent since the warping is sequential and nonlinear. If the tokens are used in a different order, a different template will result. For tokens obtained from the same speaker and spoken within the same context, order dependence is not a problem. However, for tokens obtained from different syntactic positions, order dependence is potentially a problem. Finally, if different speakers are involved, tokens will be less similar, and the order in which they are taken may greatly affect the final template. If clustering algorithms are used to generate multiple templates for each word (Rabiner, 1978), then each cluster may be viewed as a group in which order dependence may be a consideration.

3.2 Time Alignment

All but one of the parametric distance measures explored are derived from Euclidean functions of parameters pertaining to pairs of time frames. The appropriate time frames are chosen to best align the significant acoustic events in time. Because the segments aligned are monosyllabic words, one can take advantage of a number of well defined acoustic features to guide the alignment procedure. For example, the release of a prevocalic voiced stop or the onset of frication of a postvocalic fricative manifest themselves by means of such acoustic features. The particular alignment procedure used meets these requirements without requiring explicit decisions concerning the nature of the acoustic events.

The alignment operation employed a modified form of the dynamic programming algorithm first applied to spoken words by Velichko and Zagoruyko (1970) and subsequently modified by Bridle and Brown (1974) and Itakura (1975). In view of the intent to use the same algorithm for template generation as for recognition of unknown tokens, a symmetric dynamic programming algorithm was utilized. Sakoe and Chiba (1978) have recently shown that a symmetric dynamic programming algorithm yields better word recognition results than previously used asymmetric forms.

Execution of the algorithm proceeded in two stages (Figure 4). First, the pair of tokens to be compared was time aligned by appending silence to the marked endpoints and linearly shifting the shorter of the pair with respect to the longer to achieve a preliminary distance minimum. Since monosyllabic words generally possess a prominent syllabic peak in energy, this operation ensured that the syllabic peaks were lined up before the nonlinear minimization process was started. Informal evaluation has shown that use of the preliminary alignment procedure yields better results than omitting the procedure or using a linear time warping procedure to equalize the time durations of the tokens. The two tokens, extended by silence where necessary, were then subjected to the dynamic programming search to find an improved distance minimum. The preliminary distance minimum, found as a result of the initial linear time alignment procedure, corresponded to the distance computed along the diagonal of the search space and represented in most cases a good starting point for the subsequent detailed search. Use of this preliminary time alignment, and the additional invocation of a penalty function when the point selected along the dynamic programming path implied unequal time

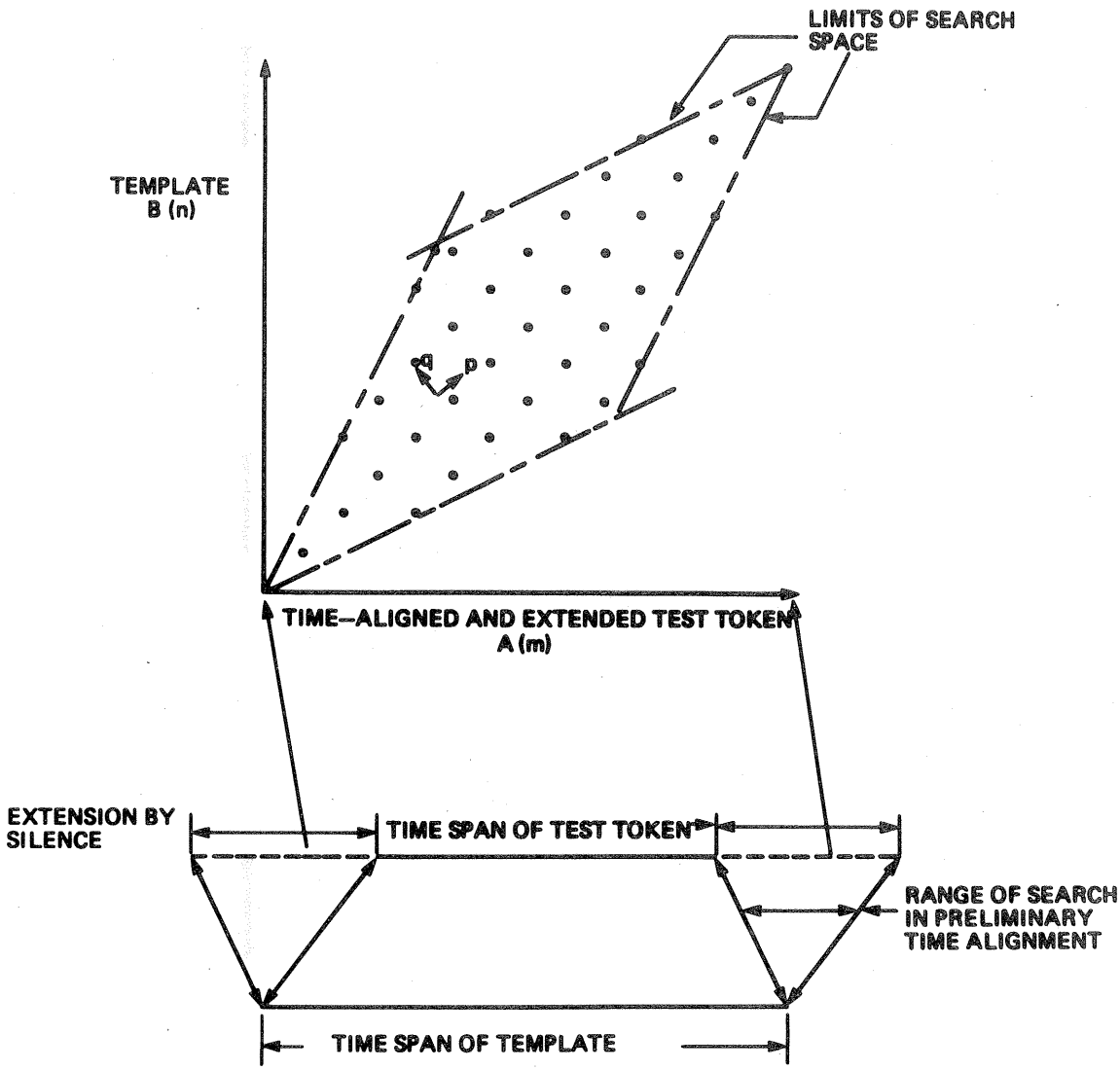


Figure 4: Dynamic time alignment of speech samples.

increments along the measured data, generally forced the optimum warping path to be near the diagonal, unless prominent acoustic information was present to indicate the contrary. For efficiency in programming, zeros (representing silence) were never really appended to the data, rather, the time shift was retained and used to trigger a modified Euclidean or Itakura distance measure when appropriate.

The use of silence to extend the syllable tokens in the preliminary time alignment, instead of linear time expansion or contraction as implied by asymmetric formulations of the dynamic programming algorithm, requires some justification. The comparison here is among syllable-sized units which generally possess an energy peak near the center regions and lesser energy near the ends. Based on a perceptual model, extension of the tokens by silence is clearly appropriate. Linear time scale changes would obscure equally the more significant duration information in the consonantal regions and less significant duration information in the vocalic regions. Discrimination between words like "pool" and "fool" depends critically on the duration of the prevocalic burst or fricative. The alignment ensures that the prominent vowel regions are lined up before time scale changes in the consonantal regions are examined.

3.3 Dynamic Warping Algorithm

The dynamic warping algorithm serves to estimate the similarity between an unknown token and a reference template. Additionally, it serves to align a reference token with a partial template to ensure that phonetically similar spectral frames are averaged in generating a composite template. Through the preliminary alignment procedure discussed above, the token or template, whichever is shorter, is extended by silence frames on both sides. The resulting multidimensional acoustic representations of the pair of patterns compared can be denoted by $A(m)$, $m = 1,2,\ldots,M$ and $B(n)$, $n = 1,2,\ldots,M$. For each pair of frames $\{A(m), B(n)\}$, a local distance function $D[A,B]$ can be defined for estimating the similarity at point $x'(m,n)$. A change of variables identifies $x'(m,n)$ as $x(p,q)$, where p and q are measured along and normal to the diagonal illustrated in Figure 4. For each position along the diagonal $\{x(p,0), 1 \leq p \leq M\}$, points along the normal $\{x(p,q), |q| \leq Q(p)\}$ are analyzed, where the search space is limited by $|q| \leq Q(p)$. The $Q(p)$ define a region in the grid area delimited by lines with slopes 1/2 and 2 passing through the corners $x(0,0)$ and $x(M,0)$.

In order for a grid point $x(p,q)$ to be an acceptable continuation of a path through some previous point $x(p-1,q')$, it must satisfy two continuity conditions:

a) $|q-q'| \leq 1$; this condition restricts the path to follow non-negative time steps along the time coordinates of the patterns, and

b) $|q-q''| \leq 1$, where $x(p-2,q'')$ is the selected predecessor of the point $x(p-1,q')$; this condition restricts any one time frame to participation in at most two local comparisons.

With the aid of these constraints, each point in the search is restricted to at most three possible predecessors. To establish the minimal distance subpath $D_T(p,q)$ leading back to the origin from the point $x(p,q)$, the cumulative distance leading to that point through each possible predecessor $x(p-1,q')$ is minimized. Thus

$$D_T(p,q) = \min_{q'} \{D_T(p-1,q') = D[A(p-q),B(p+q)]\ V(q-q')\} \quad . \qquad (9)$$

V is a penalty function introduced to keep the alignment path close to the diagonal unless a significant distance reduction is obtained by following a different path. By setting V to 1.5 for $|q-q'| = 1$ and 1.0 otherwise, unproductive searches far from the diagonal are avoided. Since all paths terminate at x(M,0), the total distance of the minimum distance path and therefore the distance between A and B is given by $D_T(M,0)$.

The minimal distance subpath passes through the points $\{x(p,\hat{q}),\ 1 \leq p \leq M\}$. These points allow the identification of pairs of frames $A(p-\hat{q})$ and $B(p+\hat{q})$ that contributed to the minimal distance result. A new template C(p), $p = 1,2,\ldots,M$, can then be generated by appropriately averaging the frames $A(p-\hat{q})$ and $B(p+\hat{q})$, $p = 1,2,\ldots,M$.

The one exception to template generation by weighted averaging occurs with the LPC. If two LPC vectors are averaged, stability of the resultant vector is not guaranteed. Therefore, LPC templates were generated in the space of LP-derived reflection coefficients. Since the reflection coefficients are bounded in magnitude by one, stability requirements are satisfied and the symmetric dynamic warping algorithm could be used without modification. Alternately, the templates could be derived in the space of LP-derived autocorrelation coefficients, since stability is guaranteed from the result that a stable autocorrelation matrix is positive definite, and a linear combination of positive definite matrices is positive definite and hence stable.

3.4 Effects of Order In Generating a Template

As discussed above, the incremental addition of individual tokens to a previously formed template results in a final template whose values depend on the order of the tokens.

In a preliminary experiment utilizing the same data base (Davis, 1979), ten sets of reference templates based on six MFCC were generated. Each set of templates used the reference tokens in random order. Independent test data were then matched with each set of templates on a per speaker basis. The average recognition scores and standard deviations were $94.76 \pm 0.53\%$ and $90.53 \pm 0.48\%$ for each speaker respectively. Thus, random ordering of tokens for template generation did not change the results. At a 0.01 significance level, none of the rates for either speaker was significantly different from the respective mean. Thirty-two of the 52 different CVC word types were never misidentified. Errors were generally confined to the same tokens of a word regardless of the template, and the most confusions were among test-reference pairs such as wake-bait, book-hood and burn-herd.

The consistent rates among template sets indicated that the templates for any given word were relatively similar. To visualize such relationships, all of the pairwise distances for eight templates and four test tokens of keep were measured and fitted to an X-Y plane. The eight templates were arbitrarily chosen from among the 24 possible templates for four reference tokens from DZ1, and the four test tokens were obtained from DZ2. The fitting procedure

was based on iterating (x,y) coordinates for test each point (template or token) until the mean-square error in distances among the points was minimized. The coordinate plane is shown in Figure 5. Regardless of ordering, the templates are close to each other and relatively far from the test tokens, thus illustrating the robustness of the technique for template generation.

4. Recognition

For each parametric representation (MFCC, LFCC, LPCC, LPC and RC), the following test procedure was used (Davis & Mermelstein, 1978). Each segmented token from sets DZ1, DZ2, LL1 and LL2 was analyzed and a matrix of coefficients (columns corresponding to coefficient number and rows corresponding to time frame) was stored (Figure 6). Each set was used in turn as test and reference data. In the case of reference data, templates were formed on a per speaker per session basis, using all tokens of each word (generally three to five in number) recorded in the session. Two types of testing were used: closed tests, where test and reference data were from the same session, e.g., reference DZ1 vs. test DZ1, and open tests, where test and reference data were from different sessions, e.g., reference DZ1 vs. test DZ2 (Figure 7). For each test word, a warping was performed with each of the 52 reference templates, and the word was identified with the least distant template (maximum similarity). In a practical situation, alternative methods, such as vowel preselection and thresholding for early rejection, could be applied to reduce the computations and the number of comparisons. In this experiment, however, the emphasis was on methodology rather than efficiency.

The results are listed in Table 1 and displayed in Figure 8 for open tests with 10 coefficients and 6.4 ms frames. Regardless of the frame separation, type of testing or speaker, these data indicate superior performance of the MFCC when compared with the other parametric representations. In fact, the performance of six MFCC was also better than any other ten coefficient set. In all cases, the 6.4 ms frame separation produced better performance. As previously stated, the window size was 25.6 ms, and using half the window size produced degraded results. Finally, speaker DZ, a male with exceptionally low fundamental frequency, was better recognized than speaker LL, a male with somewhat higher fundamental frequency. Speaker dependent differences, however, require further systematic investigation.

Most confusions arose between pairs of words that were phonetically very similar. For example, of the eight misrecognitions using the MFCC parameters for speaker DZ, two were between "bar" and "mar," two were between "pool" and "fool," one each between "keep" and "heat," "bait" and "wake," "hook" and "rig," and "hood" and "cause." Note that by not using the average spectrum energy (the zeroth cepstrum coefficient) in these comparisons, the overall energy between time aligned spectral frames has been equalized. Inclusion of the variation of overall energy with time might possibly assist discrimination between such highly confusable word pairs.

Table 1

Recognition Rates Resulting from Use of Various Acoustic Representations

Acoustic Representation	Number of Coefficients	Distance Metric	Frame Separation (ms)	Speaker	Open Test %	Closed Test %
mel-frequency cepstrum	10	Euclidean	6.4	DZ	96.5	99.4
				LL	95.0	99.1
			12.8	DZ	95.6	99.4
				LL	93.8	97.9
mel-frequency cepstrum	6	Euclidean	6.4	DZ	96.5	99.4
				LL	92.0	97.6
			12.8	DZ	95.0	98.8
				LL	90.2	97.6
linear-frequency cepstrum	10	Euclidean	6.4	DZ	94.7	99.1
				LL	87.6	98.2
			12.8	DZ	93.2	98.8
				LL	84.9	97.3
linear-prediction cepstrum	10	Euclidean	6.4	DZ	92.6	99.1
				LL	87.3	98.2
			12.8	DZ	91.7	98.2
				LL	86.4	96.7
linear-prediction spectrum	10	Itakura	6.4	DZ	85.2	97.9
				LL	84.3	95.2
reflection coefficients	10	Euclidean	6.4	DZ	83.1	97.1
				LL	77.5	97.0
			12.8	DZ	80.5	97.6
				LL	74.6	96.2

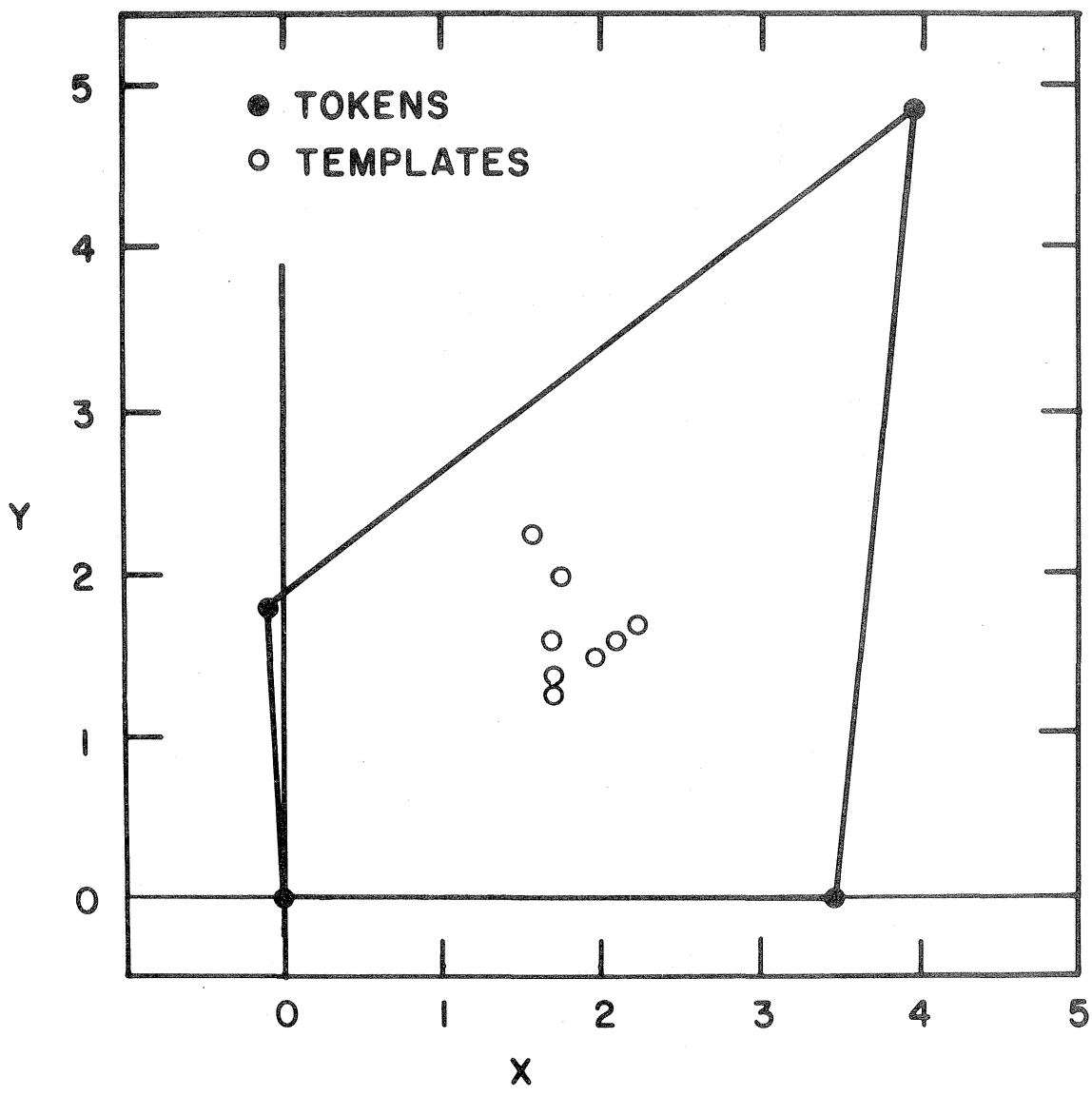


Figure 5: X-Y coordinate plane for keep.

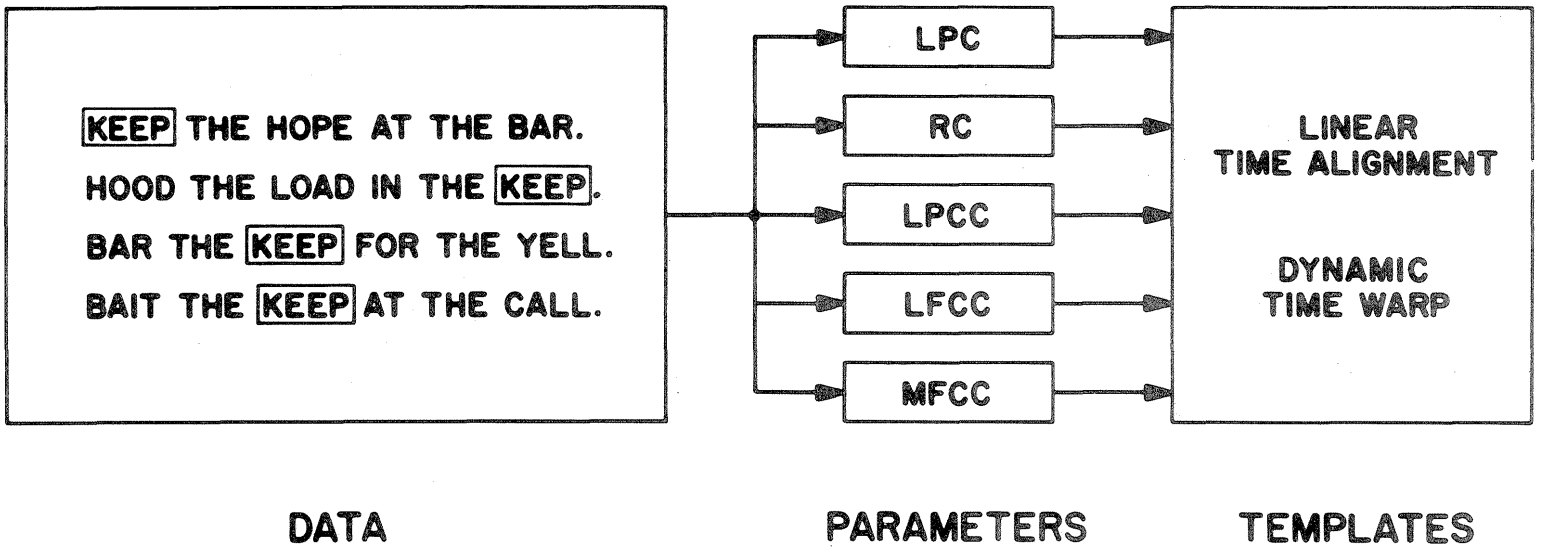


Figure 6: Selection of monosyllabic words for template generation.

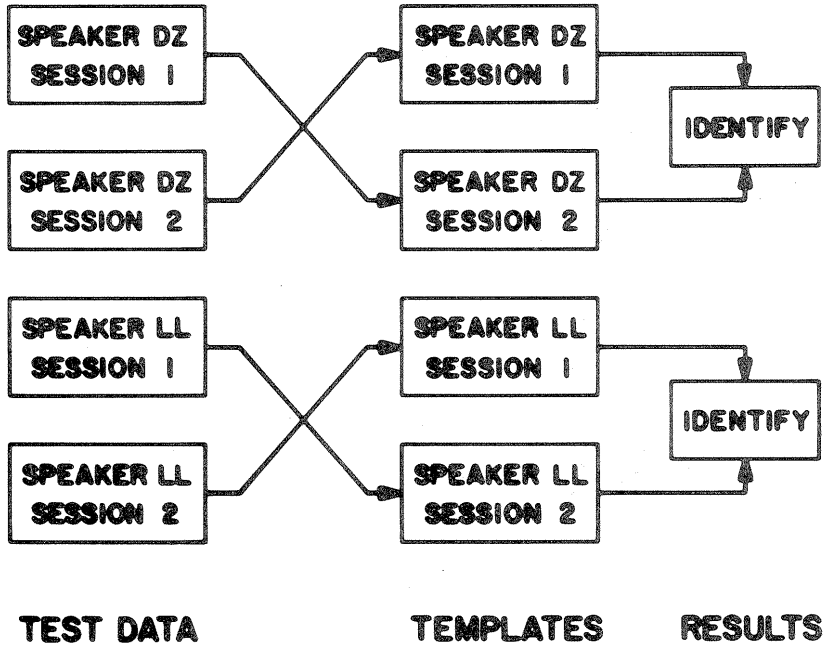
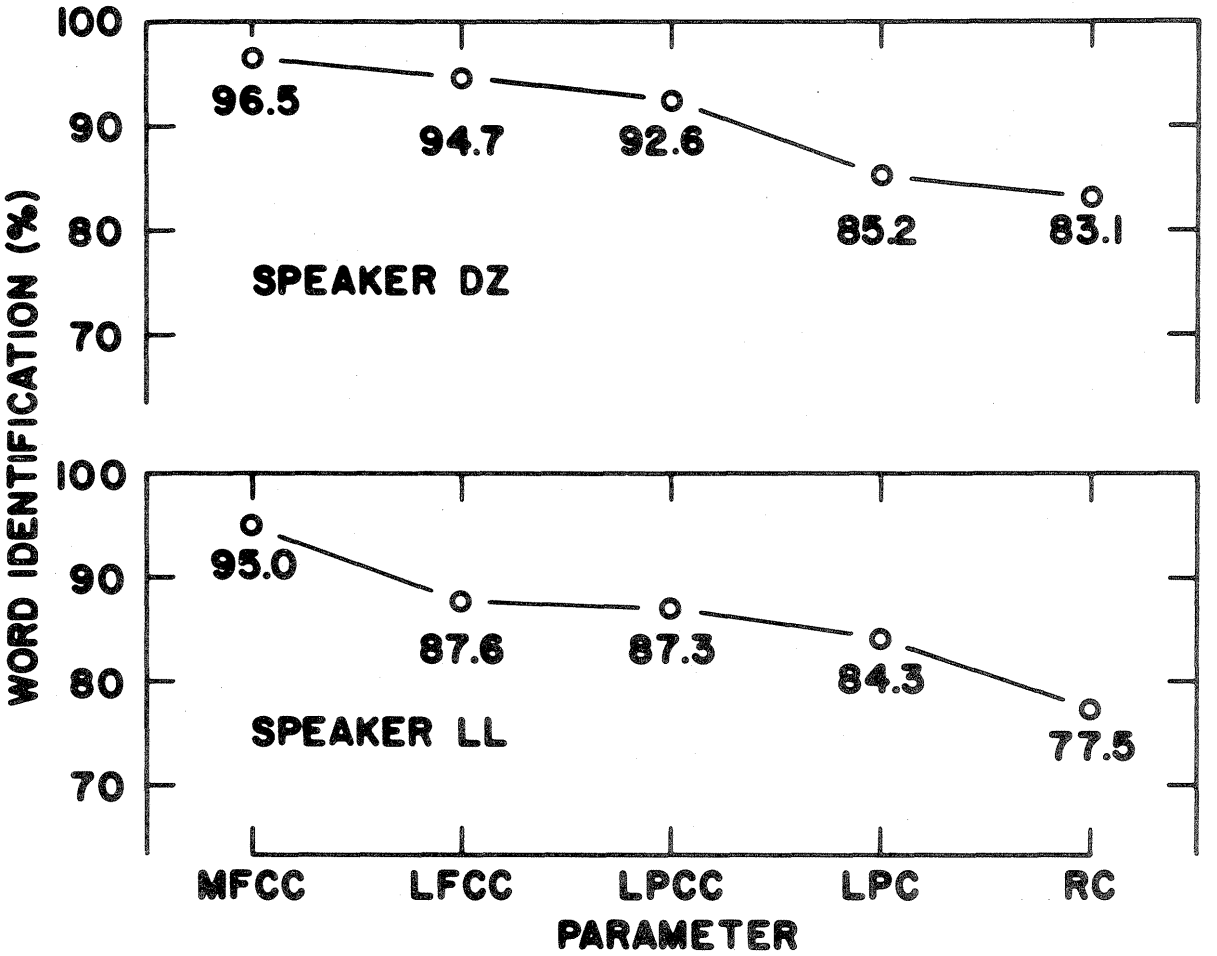


Figure 7: Two-way speaker-dependent identification tests.

Figure 8: Performance of parametric representations for recognition.



5. Conclusions

The similarity in rank order of the recogniton rates by representation for each of the two speakers suggests that the performance differences among the various acoustic representations are significant. These differences lead to the following specific conclusions:

1) Parameters derived from the short-term Fourier spectrum (MFCC, LFCC) of the acoustic signal preserve information that parameters from the LPC spectrum (LPCC, LPC, RC) omit. Both spectral representations are considered adequate for vowels. However, it is the confusions between the consonants that are most frequent. The differences found may be due to the insufficiently accurate representation of the consonantal spectra by the linear prediction technique.

2) The mel-frequency cepstra possess a significant advantage over the linear-frequency cepstra--specifically, MFCC allow better suppression of insignificant spectral variation in the higher frequency bands.

3) The cepstrum parameters (MFCC, LFCC and LPCC), which correspond to various frequency smoothed representations of the log-magnitude spectrum, succeed better than the LPC and RC in capturing the significant acoustic information. A Euclidean distance metric defined on the cepstrum parameters apparently allows a better separation of phonetically distinct spectra. Since there is a unique transformation between a set of LPCC and the corresponding LPC and RC, these representations can be said to contain equivalent information. However, this transformation is nonlinear. Representing the acoustic information in the hyperspace of cepstrum parameters favors the use of a particularly simple distance metric.

4) Defining the metric on the basis of the Itakura distance is less effective than defining it on the basis of cepstrum distance. The point of optimality is the same, i.e., equality between cepstra implies zero difference in prediction residual energy. However, the Itakura distance is less successful in indicating the phonetic significance of the difference between a pair of spectra than the cepstrum distance.

5) The mel-frequency cepstrum coefficients form a particularly compact representation. Six coefficients succeed in capturing most of the relevant information. The importance of the higher cepstrum coefficients appears to depend on the speaker. Further data are required from additional speakers before firm conclusions can be reached on the optimal number of coefficients.

The results are limited by the restrictions on the speech data examined. In particular, consonant clusters, multisyllabic words and unstressed monosyllabic words have not been studied. Expansion of the data base along any one of these directions introduces additional representation problems. It is not obvious that the best representation for stressed words is also best for the much more elastic unstressed words. These questions are left for future studies.

It should be emphasized that the comparative ranking of the representations can be influenced by the choice of both the local and the integrated distance metrics. A Euclidean distance function is one of the simplest to implement. However, taking into account the probability distributions of the individual parameters should result in improved performance. Estimating these distributions requires considerable data. Yet, even if only a few parameters of these distributions are known, for example, the variance of the cepstrum coefficients, better local distance metrics could be designed. Despite the high recognition rates achieved so far, there is reason to believe that even better performance can be attained in the future.

The design of the mel-frequency cepstrum representation was motivated by perceptual factors. Evidently, an ability to capture the perceptually relevant information is an important advantage. The design of an improved distance metric may result from more accurate modeling of perceptual behavior. In particular, where a constant difference between spectra persists for a number of consecutive time frames, the contribution of that difference in the current distance computation is proportional to the duration of that difference. With the possible exception of very short durations, no perceptual justification exists for this property (Feldtkeller & Zwicker, 1956). Nevertheless, the distance function must in some fashion combine different information from all the time frames constituting the signals compared. Further optimization of the integrated distance function represents an important challenge.

For each representation a small but significant gain in recognition is achieved by decreasing the frame spacing from 12.8 ms to 6.4 ms samples. The average difference in the recognition rates is 1.7%. However, the computational complexity for any dynamic programming comparison varies as the square of the average number of frames constituting a word. Thus a significant computational penalty accompanies any increase in the frame rate. In contrast, the computations grow only linearly with the number of cepstrum coefficients. Since the recognition rates for six cepstrum coefficients and 6.4 ms frame spacing is quite comparable to the rate for ten coefficients and 12.8 ms frame spacing, increasing the number of coefficients and maintaining a somewhat coarser time resolution is computationally more advantageous than using fewer coefficients more frequently.

The principal conclusion of the study is that perceptually based word templates are effective in capturing the acoustic information required to recognize these words in continuous speech. Due to the various limitations of this study, a conclusion that such high recognition rates are attainable with a complete automatic system operating in a practical environment is not warranted at this time. However, the results do encourage a continuing effort to optimize the performance of speech recognition systems by critical evaluation of each of the constituent components.

REFERENCES

Bridle, J. S., & Brown, M. D. An experimental automatic word recognition system. JSRU Report (Joint Speech Research Unit, Ruislip, England), 1974, No. 1003.

Davis, S. B. Order dependence in templates for monosyllabic word identifica-

tion. Conference Record, 1979 International Conference on Acoustics, Speech and Signal Processing, Washington, 1979, 570-573.

Davis, S. B., & Mermelstein, P. Evaluation of acoustic parameters for monosyllabic word identification. Journal of the Acoustical Society of America, 1978, 64, Suppl. 1, S180. (Abstract)

Fant, C. G. M. Acoustic description and classification of phonetic units. Ericsson Technics, 1959, 1. Also in G. Fant, Speech sounds and features, MIT Press, 32-83, 1973.

Feldtkeller, R., & Zwicker, E. Das Ohr als Nachrichtenempfanger. Stuttgart: S. Hirzel, 1956.

Fujimura, O. The syllable as a unit of speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1975, ASSP-23, 82-87.

Gray, A. H. Jr., & Markel, J. D. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech and Signal Processing, 1976, ASSP-24, 380-391.

Itakura, F. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1975, ASSP-23, 67-72.

Markel, J. D., & Gray, A. H. Jr. Linear prediction of speech. New York: Springer-Verlag, 1976.

Mermelstein, P. Automatic segmentation of speech into syllabic units. Journal of the Acoustical Society of America, 1975, 58, 880-883. (a)

Mermelstein, P. A phonetic-context controlled strategy for segmentation and phonetic labelling of speech. IEEE Transactions on Acoustics, Speech and Signal Processing, 1975, ASSP-23, 79-82. (b)

Mermelstein, P. Distance measures for speech recognition, psychological and instrumental. In C. H. Chen (Ed.), Pattern recognition and artificial intelligence. New York: Academic Press, 1976, 374-388.

Mermelstein, P. Recognition of monosyllabic words in continuous sentences using composite word templates. Conference Record, 1978 International Conference on Acoustics, Speech and Signal Processing, Tulsa, 1978, 708-711.

Pfeifer, L. L. Interactive laboratory system users guide. Santa Barbara: Signal Technology, Inc., 1977.

Pols, L. C. W. Spectral analysis and identification of Dutch vowels in monosyllabic words. Unpublished doctoral dissertation, Free University, Amsterdam, 1977.

Rabiner, L. R. On creating reference templates for speaker independent recognition of isolated words. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, ASSP-26, 34-42.

Rabiner, L. R., Rosenberg, A. E., & Levinson, S. E. Considerations in dynamic time warping algorithms for discrete word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, ASSP-26, 575-586.

Sakoe, H., & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, ASSP-26, 43-49.

Schroeder, M. R. Recognition of complex acoustic signals. Life Sciences Research Report, T. H. Bullock ed., 1977, 55, 323-328.

Velichko, V. M., & Zagoruyko, N. G. Automatic recognition of 200 words. International Journal of Man-Machine Studies, 1970, 2, 223-234.

Wakita, H. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Transactions on Acoustics, Speech and Signal Processing, 1973, AU-21, 417-427.

White, G. M., & Neely, R. B. Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming. IEEE Transactions on Acoustics, Speech and Signal Processing, 1976, ASSP-24, 173-188.

FOOTNOTE

[1]Fant (1973) compares Beranek's mel-frequency scale, Koenig's scale and Fant's approximation to the mel-frequency scale. Since the differences between these scales are not significant here, the mel-frequency scale should be understood as a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

APPENDIX A

Sentences used for word recognition.

1. Keep the hope at the bar.
2. Dig this rock in the heat.
3. Wake the herd at the head.
4. Check the lock on the seal.
5. Bang this bar on the head.
6. Call a mess in the case.
7. Cut the coat for a mop.
8. Foot the work in the mess.
9. Boot the back of the book.
10. Burn your check in the jar.
11. Mop the room on the watch.
12. Load the tar for the bait.
13. Tar this rig in a rush.
14. Fear a hood on the ship.
15. Rig a bait for the work.
16. Nail that book to the rock.
17. Yell this call for the wake.
18. Gang the bait on the coat.
19. Walk the watch in the hope.
20. Buff one book for the walk.
21. Hook the mop on the lock.
22. Pool the case for the man.
23. Hurl his bar in the muck.
24. Bomb the head at the wake.
25. Pose this seal for the gang.
26. Mar the watch on the hood.
27. Heat the foot of the fool.
28. Kill the herd for the load.
29. Case your ship for the cause.
30. Head the rush for the burn.
31. Back the pool for the check.
32. Watch that hook with the nail.
33. Rush the buff at the foot.
34. Hood the load for the keep.
35. Room one seal in the pool.
36. Herd the fool with a yell.
37. Rock the mop with a hurl.
38. Coat the cut with the tar.
39. Jar the bomb with a bang.
40. Seal the dig in a fear.
41. Ship the nail in a boot.
42. Bait the keep with a call.
43. Mess his work in the room.
44. Man the cut at the kill.
45. Cause a mar on the back.
46. Muck the gang on the walk.
47. Book the fool on the rig.
48. Fool the man on the rock.
49. Work the hurl at the dig.
50. Lock your man in a pose.
51. Hope this call for the heat.
52. Bar the keep for the yell.
53. Put a bang in the bomb.
54. Set a pose in the muck.
55. Pose a jar on the buff.
56. Kill the fear in the cause.
57. Mar the burn on the head.