Canadian Acoustics / Acoustique canadienne, 1989, 17 (3): 3-13

RECOGNITION OF THE SPOKEN FRENCH ALPHABET USING A TWO-PASS DYNAMIC TIME WARP ALGORITHM

J-P. Cordeau and P. Mermelstein
INRS - Télécommunications
3, place du Commerce
Ile-des-Soeurs
Verdun,Québec
H3E 1H6

765-9999

ABSTRACT

A dynamic time warp algorithm for isolated word recognition is adequate when the vocabulary consists of easily discriminable words. On the other hand, when the words are taken from a vocabulary whose elements are acoustically similar, specialized algorithms must be used to help the recognition task. The spoken French alphabet is composed of acoustically similar words that can be organized into 6 distinct classes. For recognizers working on such a data base, the words not recognized properly are always within-class words and a special recognition approach must be introduced to overcome these effects. In this paper, a two-pass algorithm is used to discriminate between members of the highly confusable sets [1]. The first pass is used to differentiate similar word classes, while the second pass uses a discriminant analysis to reliably separate within-class tokens. The second stage provides better discrimination, separating the words within each class through improved focus on the temporal and cepstral regions of greatest between-word variance. The improvement in discriminability is provided by a modified frame-specific weighting scheme. The error rate is reduced by 50 % (from 40 % to 20 % considering only the letters which cause some errors in the first pass) using this approach as compared to the direct one-step DTW algorithms.

SOMMAIRE

La reconnaissance automatique de mots isolés par anamorphose temporelle (ou DTW) est un outil adéquat lorsque les mots dictés se distinguent facilement les uns des autres. Par contre, lorsque ces mots sont tirés d'une base de données à haut niveau de difficulté, où les mots sont acoustiquement semblables, un algorithme spécifique doit être employé pour faciliter la tâche. L'alphabet français est un vocabulaire qui se compose de lettres similaires que l'on peut regrouper en 6 classes distinctes. Les erreurs obtenues par les sytèmes de reconnaissance travaillant sur cette base de données sont uniquement dues à des fautes intra-classes. Pour pallier à cette carence, un système à deux étapes est introduit. La première étape est utilisée pour différencier les classes tandis que la seconde procède à une discrimination intra-classe, favorisant la séparation des lettres acoustiquement semblables. La deuxième étape améliore la discrimiation en pondérant davantage les coefficients utiles à la dissociation (les régions temporelle et cepstrale de grande variance) en modifiant légèrement la notion de distance entre la lettre test et la lettre de référence. Le taux d'erreur est diminué de moitié (de 40 % à 20 % en ne considérant que les lettres similaires) en utilisant cette méthode comparativement à une méthode d'anamorphose simple.

I - INTRODUCTION

The theoretical problem for isolated word recognition (or IWR) is to discover the invariants in speech which occur when the same words are spoken by different (in speaker independant systems) or identical speakers (in speaker dependent systems). Although these variations are seldom perceived by human listeners, they can be seen by either a frequency or a time domain analysis. For two identical words spoken by the same speaker, there will be important acoustical changes in the waveforms due to the emotional state of the talker: one of the versions will be longer; the other will be louder; etc. An automatic recognition system must therefore compensate for these anomalies by focusing on the underlying phonetics of the words which are constant, and by neglecting other levels of information, wheter prosodic or aesthetic.

Dynamic time warping (DTW) algorithms have been used extensively for IWR and have undergone some refinement [2], [3]. In most cases, the improvements serve to speed up the algorithm or to take into account various problems due to the segmentation of word templates (constrained end point warp). DTW is based on four underlying principles [4]: 1) global variations are adequately treated by a nearly linear warp; 2) local variations can be treated through a dynamic approach where weights are used when the path deviates from its linear course; 3) each frame of the test utterance has the same importance and can be considered independent and 4) a uniquely defined distance measure is sufficient for the comparison of words in the search space. These four considerations have led to a multitude of straightforward pattern similarity measures which differ at some point in the comparison of the test and reference patterns (initialisation step, the dynamic approach or the warping techniques). However, all share the same problem which is related to the third and fourth assumptions cited above: the versatility of the DTW algorithms introduce difficulties for highly complex vocabularies.

A practical isolated or continuous word recognizer must have the flexibility to accept spelled words so as to allow a missclassification to be spelled out and be correctly recognized. This paper deals with the recognition of the spoken French alphabet and the accented letter *ET* (é), a highly confusable set. It is composed of one-syllable words (except for *W* and *Y*) often formed by a consonant and a vowel. The misclassified words can be organized into six distinct classes, each of them composed of acoustically similar letters, the other letters not included in $\Phi(1)$ to $\Phi(6)$ are unambiguous and are not included in the classes:

$$\Phi(1) = \{A, K\} = \{\ /\text{a}/,/\text{ka}/\ \}$$
$$\Phi(2) = \{ET, B, C, D, G, P, T, V\}$$
$$= \{\ /\text{e}/,/\text{be}/,/\text{se}/,/\text{de}/,/\text{ʒe}/,/\text{pe}/,/\text{te}/,/\text{ve}/\ \}$$
$$\Phi(3) = \{I, J\} = \{\ /\text{i}/,/\text{ʒi}/\ \}$$
$$\Phi(4) = \{U, Q\} = \{\ /\text{y}/,/\text{ky}/\ \}$$
$$\Phi(5) = \{F, S\} = \{\ /\text{ɛf}/,/\text{ɛs}/\ \}$$
$$\Phi(6) = \{M, N\} = \{\ /\text{ɛm}/,/\text{ɛn}/\ \}$$

The phonetic transcription of the letters lead to two considerations: 1) the differences between the classes are mainly vocalic and 2) the within class differences come from adding or changing a consonant. Although isolated word recognizers have no problem distinguishing letters between classes, they all have difficulty discriminating between letters of the same class. A typical recognizer working on the comlete alphabet will have a 90 % recognition score, which will fall abruptly to 60 % considering only the six classes mentioned above.

The problem is the inability of the DTW to take into account the discriminating factors localized in time and frequency which differentiate one letter from another. The following paragraphs will consider only the second class, also known as the E-set, which accounts for most of the misclassifications.

The E-set has 8 letters composed of either a fricative followed by the vowel /e/ (C,G,V); a stop followed by the same vowel (B,D,P,T); or simply the vowel alone (for the accented letter ET or é). Since DTW considers all parts of the utterance to be equally important, the region that represents the initial portion of the words, the only region which can distinguish them appart, has the same weight as the region representing the identical /e/ vowel at the end of the letter. Moreover, the vowel is responsible for more than 3/4 of the total duration of the letter, which further complicates the recognition task.

Besides the temporal considerations stated above, other time factors associated with the consonant tend to augment the misclassification of letters. The French plosives can be decomposed into 4 sections:
• a **silence** which corresponds the the total obstruction of the vocal tract. For the voiced stops the silence is accompanied by a voice bar, i.e energy in the low frequencies (100-300 Hz) that corresponds to glottal radiation.
• The **explosion** when the vocal tract opens which gives out a short burst of energy at frequencies which depend on the place of articulation.
• a **friction noise** coming from the turbulence near the obstruction and having a spectra similar to the one of the noise.
• **formant transitions** to the next phoneme.

The characteristics which distinguish the voiced stops from their unvoiced conterparts lie in the presence/absence of the voice bar; the duration of the friction noise; and the length of the voice onset time (or VOT). The latter corresponds to the time delay from the explosion to the begining of voicing (an increase in energy at all frequencies [5]). For the French stops /b/, /d/and /g/, the voicing precedes the burst and the VOT is thus assumed negative, whereas the plosives /p/, /t/and /k/are attributed positive VOT. Information about place of articulation (labial, alveolar and velar for /b//p/, /d//t/and /g//k/respectively) is found in the spectrum, although the burst duration is proportional to the inertia of the articulator (lips tend to move faster than the back of the tongue). For labial stops, the burst is weak since no pressure buildup is possible after the obstruction, and it is hard if not impossible to locate acurately in the frequency domain. For alveolar stops followed by the vowel /e/, it can be affirmed that the burst is higher in frequency for /t/ than for /d/, the former being located near 4 KHz and the latter near 3.6 KHz [6]. Moreover, both of the alveolar burst frequencies are higher than those of the velars ($\simeq$ 2.8 KHz). Studies have also confirmed that important discriminating factors are found in the formant transitions, from the plosive to the ensuing vowel. It then seems that static and dynamic features are both responsible for the correct recognition of all the stops found in the French phonetics [7] [8].

DTW recognizers cannot apply these concepts in a straightforward manner since the distance measure integrates, in the local difference between the reference and test frame, all the frequency information on a uniform basis. Furthermore, as we have seen, the discriminability is not uniform in time but rather concentrated in regions which correspond to phonetically important aspects of the sound.

An approach that permits one to focus on the spectral and temporal discriminating regions will be explained in section III. Section II will present the baseline recognizer, emphasizing its inability to

take advantage of the discriminative features. Section IV will evaluate the data base, elucidate the parameters used in the front end of the system and explain how to obtain the final recognition score, while section V will summarize the results.

II - DTW RECOGNIZER

All DTW algorithms work in the same fashion: a dilation/contraction of the time scale so that similar acoustic segments of the test speech and reference templates may correspond. If R and T are the parametric representations of two utterances, they can be described by a temporally ordered time index equal to the window length of the preprocessing stage and two indices I and J equal to the total duration of the letters:

$$R = R(1), R(2), ..., R(i), ..., R(I)$$
$$T = T(1), T(2), ..., T(j), ..., T(J)$$

where every R(i), T(j) is a feature vector in the cepstral parametric space (see section IV), C_i being the i^{th} cepstral coefficient out of a total of $Ncep$ used:

$$R(i) = \begin{pmatrix} C_1^{R(i)} \\ \cdots \\ \cdots \\ C_{Ncep}^{R(i)} \end{pmatrix}, T(j) = \begin{pmatrix} C_1^{T(j)} \\ \cdots \\ \cdots \\ C_{Ncep}^{T(j)} \end{pmatrix}$$

The time warp consists in eliminating the temporal differences $|I - J|$ in an optimal fashion. To accomplish this task we must define a distance measure between the reference and test frames. In this paper, the Euclidian distance is used for its simplicity and good performance [9]:

$$d_e(T(j), R(i)) = \sum_{l=1}^{Ncep} (C_l^{T(j)} - C_l^{R(i)})^2 \tag{1}$$

The alignment procedure finds an optimal path $C = \{c(k)\} = \{i(k), j(k)\}$ through the $< i - j >$ space $(0 < i < I; 0 < j < J)$ and computes the total distance score between the test word T and every reference word R by:

$$D(T, R) = \frac{\sum_k d_e(T(j(k)), R(i(k)))p(k)}{\sum_k p(k)} \tag{2}$$

where p(k) corresponds to the warping function defined by Sakoe [3].

$$p(k) = (i(k) - i(k-1)) + (j(k) - j(k-1)) \tag{3}$$

This function is chosen to constrain the path to follow a logical course, to oblige the words to be aligned in a sensible way (for example by disallowing a too steep or a too gentle warp). Equation (1) shows how a distance score is computed at every frame between the test and reference tokens. The metric considers each coefficient to be equally important; no part of the parametric space is emphasized more than another. Equation (2) shows how the individual distance scores are accumulated, treating every frame in an equivalent manner.

III - THE DISCRIMINATION FUNCTIONS

If we plot the local distance $d_\epsilon(T(j(k)), R(i(k)))$ as a function of the warped time, we notice three types of behavior depending on the letters compared. The solid curve in Figure 1 shows the local distance between two pronunciations of the letter A. It shows a more or less uniformaly distributed function. The dashed curve is a comparison between letters of different classes (A and O). The local distance score is also uniformaly distributed, but in this case, the function is larger for all frames than the within-letter function, thus showing dissimilarity for every frame. The last behavior reflects the local distance when letters of the same class are used as test and reference tokens (K and A respectively). We see that the dissimilarity is large only for the initial portion of the word, corresponding, presumably, to the presence vs. absence of the velar stop /k/.

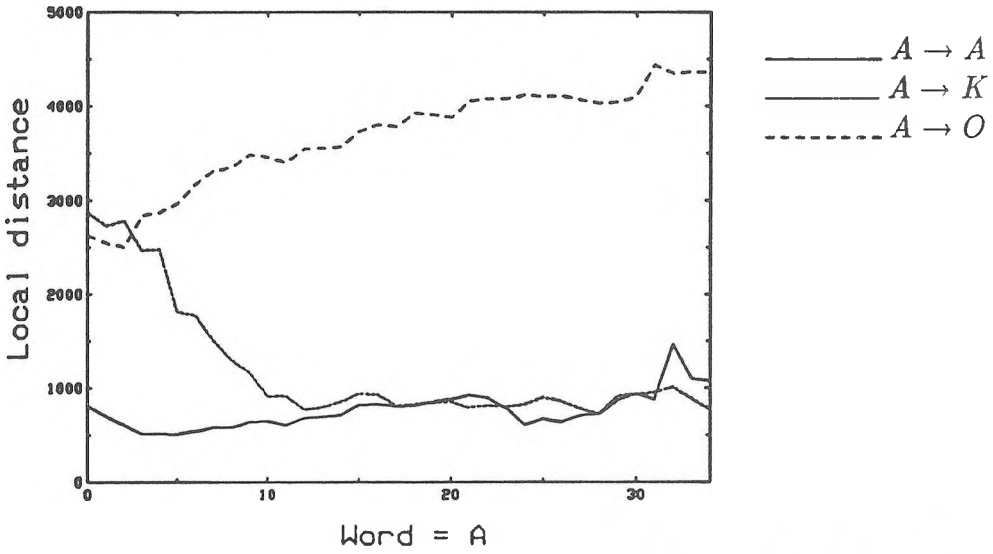


Fig. 1 Local distance for letters A,K and O. $A \rightarrow A$ (solid), $A \rightarrow K$ (dotted) and $A - O$ (dashed).

Rabiner and Wilpon [1] uses this information to create temporal weights to encourage discrimination when the local distance depart from the within-letter scores (when the dashed and dotted curves are greater than the solid curve in Figure 1). To accomplish this task, they compute weight functions $w^t(k)$ for all within-class letters and integrate them in the computation of the local distance in the second stage to improve temporal discrimination:

$$w^t(k) = \frac{|\hat{d}_2^{RR'}(k) - \hat{d}_1^{RR}(k)|}{\sqrt{\sigma_{\hat{d}_1}^2 + \sigma_{\hat{d}_2}^2}} \tag{4}$$

where d_1 and d_2 correspond to the means of local distances for within-letter and different but within-class letter respectively.

The main problem with such a scheme is that not all the letters in the French alphabet reveal the three different behaviors depicted in Figure 1. In some cases, when the normal speech variations

are high for identical letters, there will be only two distinct behaviors and the local distance function will never be uniformaly distributed as in the solid curve. This is the case for most of the letters in the E-set: d_e will reveal a large pulse for identical and different letters at the explosion and transition frames (Figure 2). For this reason, a discrimination function showing a behavior similar to that to the curves cannot be employed in a straightforward manner.

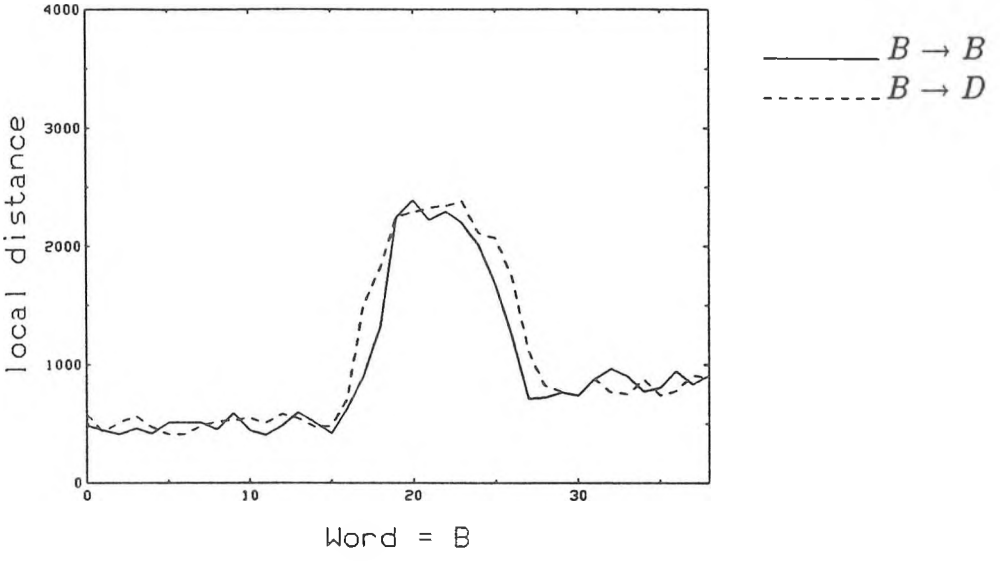


Fig. 2 Local distance for letters B and D. $B \rightarrow B$ (solid), $B \rightarrow D$ (dashed).

Temporal indices are not enough to discriminate acurately between all the words in the six classes mentioned earlier: R_1 and R_2 will be very close to one another for the burst and transition frames of the stops. The probability of correct recognition will then approach 50 % for the whole letter since the ending vowel is also the same. To overcome this effect, non uniform weighting functions should depend not only on local differences distributed across time, but also on differences defined in the cepstral domain. To accomplish this cepstral discrimination, we can enlarge the scope of equation (4) by considering individual cepstral coefficients to be normally distributed and then obtain the cepstral weighting function as:

$$w_l^c(k) = \frac{|\hat{d}_2^{RR'}(k,l) - \hat{d}_1^{RR}(k,l)|}{\sqrt{\sigma_{d_1}^2 + \sigma_{d_2}^2}} \tag{5}$$

where $\hat{d}_2$ and $\hat{d}_1$ correspond to means of cepstral differences for different and identical letters respectively:

$$\hat{d}_1^{RR}(k,l) = \frac{1}{N_{d1}^2} \sum_R \sum_{R'=R} (C_l^{R(k)} - C_l^{R'(k)})^2 \tag{6}$$

$$\hat{d}_2^{RR'}(k,l) = \frac{1}{N_{d1} N_{d2}} \sum_R \sum_{R' \neq R} (C_l^{R(k)} - C_l^{R'(k)})^2 \tag{7}$$

The weight function may then be used in the distance metric discussed earlier to emphasize cepstral regions of interest:

$$d_e^c(T,R) = \sum_{l=1}^{Ncep} w_l^c(k)(C_l^{T(k)} - C_l^{R(k)})^2 \tag{8}$$

The functions $w_l^c(k)$ depend on two distinct groups of letters R and R' (N_{d1} pronunciations of the letter R and N_{d2} pronunciation of the letter R'). They represent the weighting curves which discriminate in the temporal and cepstral space the two letters R and R'. A high amplitude for a certain coefficient at a certain frame will mean increased discriminability in this region of the cepstral-time domain. Their computation can be made in many ways. For the temporal weights w^t, Rabiner and Wilpon uses a pre-recognition stage where all pairs of within-class tokens are employed. The problem with such a scheme is that the letters are not already aligned and the means $\hat{d}_1$ and $\hat{d}_2$ do not reflect the true averages. The problem is greater when considering individual cepstral coefficients since there is no frame averaging in equations (6) and (7). To overcome these effects, an alignment procedure is interposed between the first recognition and the second discrimination stage. It relies on the identification of the burst, and the alignment of the words prior and after the explosion for all the letters in the E-set that have a stop as their first phoneme. Once the tokens are aligned, the averages $\hat{d}_1$ and $\hat{d}_2$ are computed and w_l^c can be calculated.

IV - METHOD

The vocabulary used for testing consisted of the 26 letters of the French alphabet and the accented letter *é*. The speech data was obtained by 12 repetitions of each word by a male speaker. In all, 324 files were recorded and digitized at 16 KHz. Half of the files were used as reference templates and half as test data. A manual segmentation using a high resolution screen in the spectral and temporal domain was used to separate the letters from each other. The preprocessing stage consisted in obtaining $Ncep = 7$ mel frequency cepstral coefficient (MFCC) using an analysis window of 6.4 msec and a window length of 25.6 msec [10](C_0 was not used in the recognition stage).

The first pass consists of a straightforward recognition by a DTW algorithm. To take into account errors at the segmentation stage, an unconstrained end point algorithm is used, allowing every word a 38.4 msec (6 frames) translation. Rabiner and Wilpon uses a normalization process, representing the test and reference letters by a fixed number of temporally normalized frames. Although this process facilitates the recognition, it is done at the expense of destroying important frames useful for the second pass. All the reference letters which had a total score below a given threshold, here taken as the mean of total distance scores for identical letters, were passed to the alignment and second stage. If all the letters that are passed to the second stage are identical, the word is immediately recognized and no discrimination occurs. On the other hand, if the subset of reference tokens contain different letters, the second stage is used as shown in figure 3. In this example, 10 letters are passed to the second stage: three B, three D' and four V. The test word T is thus assumed to be in one of these groups (i.e. $T \in \{B, D, V\}$). The word of each group that has the lowest distance score after the first pass is printed in boldface and is used in the second pattern similarity measure. The weights of Eq. (5) are obtained after the first recognition, considering only the letters which are close to the test word T (in the example above, 10 letters are used). The number of tokens that take part in the computation of Eq. (6) and (5) ($Nd1$ and $Nd2$) depend on the number of letters passed to the second stage.

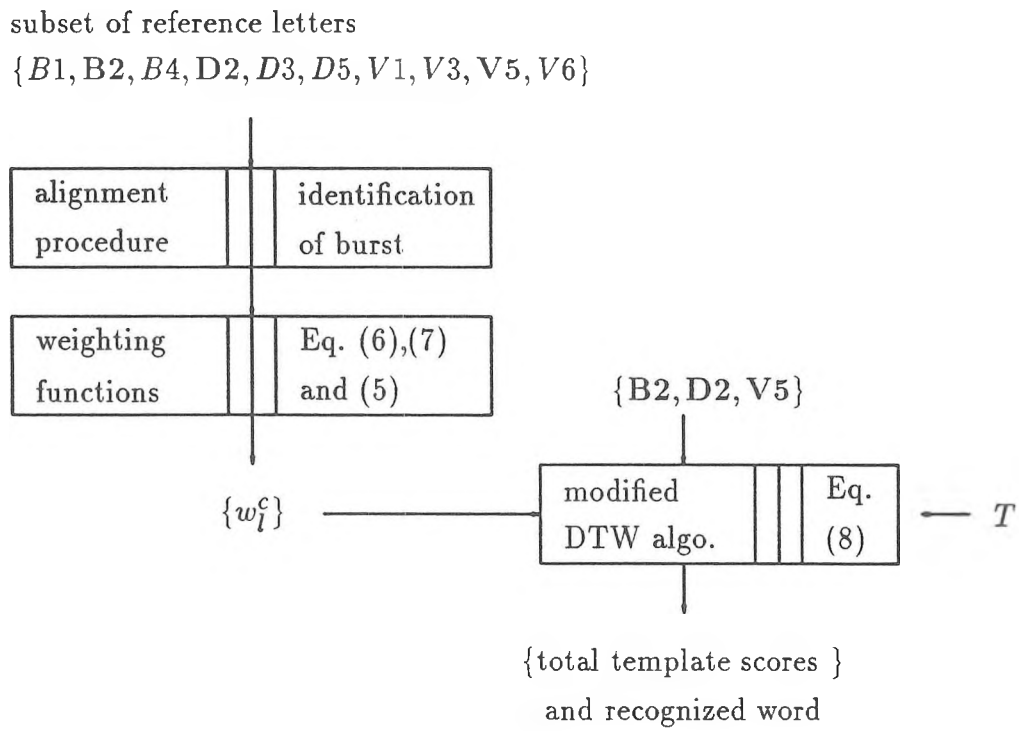


Fig. 3 The second stage of the recognizer.

V - RESULTS

The representation in figure 4 shows the cepstral weighting curves obtained from the two letters B and D ($R' = D, R = B$ in equations (6) and (7)). The abscissa "x" represents the cepstral coefficient index, the "y" axis the frame index and "z" the amplitude of the functions. Firstly, we notice that the high amplitude regions are concentrated in time, that the maximum of discrimination happens at the 19^{th} frame, where the burst takes place. The curves fall off abruptly to zero because of the alignment of the tokens. Secondly, the plot shows that only the C_1, C_2, C_3 and C_4 coefficients are important for discrimination, the others have weak amplitudes for all the frames in the word. This signifies that there is a need for cepstral discrimination since the temporal curve w^t would integrate at every frame the information of all the MFCCs into one factor.

The confusion matrices before and after discrimination are given in Tables 1 and 2.

They show that the cepstral discrimination achieves a 47 % reduction of errors when only the letters having some misclassification are taken into account. When only temporal discrimination is applied, with the w^t functions, the number of errors drop from 19 to 16, bringing only a 15 % reduction in the error rate. The total recognition when the 162 test words are used passes from 88 % to 94 %. The discriminant analysis not only helps the recognition, but also clusters some letters together: the matrix found in table 2 has all its elements very close to the diagonal which is not the case with no discrimination. This fact is beneficial since it reduces the number of non-empty cells thus reducing the number of posibilities in the recognition process.

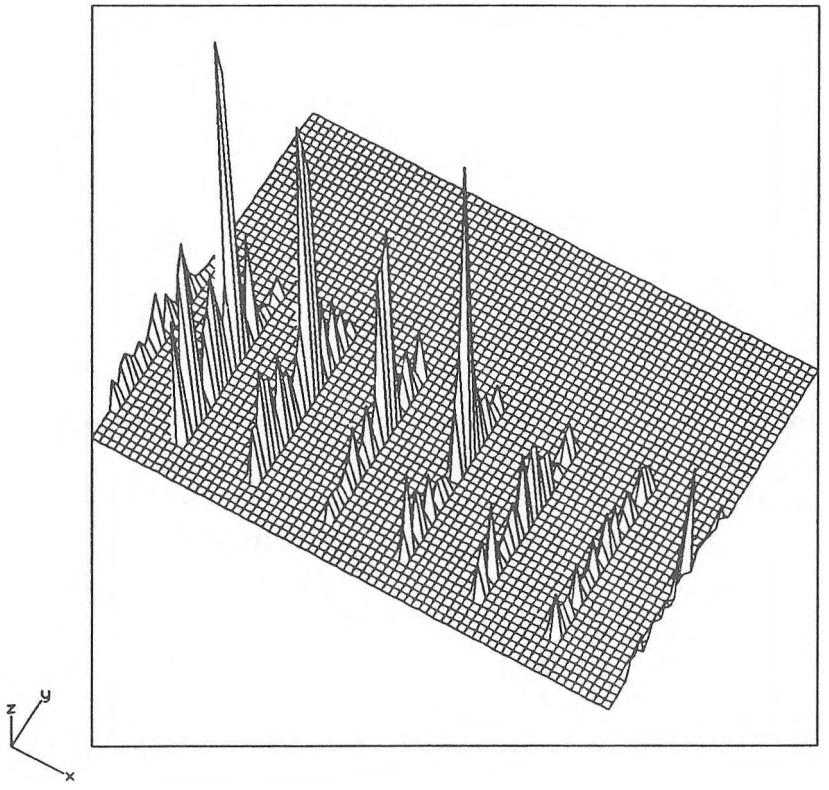


Fig. 4 Cepstral weighting curves for letters B and D in the cepstral domain; "x" is the cepstral coefficient index, "y" the frame (temporal) index and "z" the amplitude of the weight functions.

	F	S	B	D	V	P	T	ć	TO
F	6	-							6
S	1	5							6
B			6	-	-	-	-	-	6
D			5	1	-	-	-	-	6
V			1	1	4	-	-	-	6
P			-	-	-	1	4	1	6
T			1	-	-	2	1	2	6
ć			-	-	-	1	-	5	6
ER	0	1	0	5	2	5	5	1	19

Table 1 Confusion matrix without discrimination.

VI - DISCUSSION

We have shown that cepstral discrimination can improve the recognition of isolated words when they are taken from a phonetically similar vocabulary. The speaker dependant, two pass system described briefly in this paper is superior by 50 % to the straightforward DTW recognizer. Although the system as it stands cannot be used as a speaker independant system (since the weights are very speaker sensitive), by producing reference templates that come from different speakers, the system can be used

	B	D	V	P	T	ℓ	TO
B	5	1	-	-	-	-	6
D	3	2	1	-	-	-	6
V	-	-	6	-	-	-	6
P	-	-	-	2	4	-	6
T	-	-	-	1	5	-	6
ℓ	-	-	-	-	-	6	6
ER	1	4	0	4	1	0	10

Table 2 Confusion matrix with alignment and cepstral discrimination.

by the different speakers that took part in the making of the templates since the computation of the weights are done after the first recognition stage, taking into consideration only the tokens similar the test letter. The algorithm described can be placed in parallel with another recognizer having a larger lexicon, where the unrecognized test tokens have to be spelled out. This would be done at the expense of increased perplexity and slower recognition. Although the two-pass pattern recognition approach is useful, some errors still occur: out of six utterances of the letter P, four were recognized as T, showing that further improvement is still needed. Changing the preprocessing stage, having either more cepstral coefficients or finer frequency parameters more suitable to the second stage analysis (critical band filter outputs for example) would help the accuracy of the process, but again at the expense of a larger computational load and an increase in memory requirements.

REFERENCES

[1] L. Rabiner and J. Wilpon, "A two-pass pattern-recognition approach to isolated word recognition," *Bell Syst. Tech. J.*, vol. 50,No.5, pp. 739-766, May 1981.

[2] L. Rabiner et al, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust.. Speech and Signal Processing.*, vol. ASSP-26,No.6, pp. 575-582, Dec. 1978.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust.. Speech and Signal Processing.*, vol. ASSP-26,No.1, pp. 43-49, Feb. 1978.

[4] D. O'Shaughnessy, *Speech Communication.* Addison-Wesley:Boston, 1987.

[5] C. Darwin and J. Seton, "Perceptual cues to the onset of voiced excitation in aspirated initial stops," *J. Acoust. Soc. Am.*, vol. 74(4), pp. 1126-1135, Oct. 1983.

[6] V.Zue, "Acoustic characteristics of stop consonants: a controlled study," Indiana University Linguistics Club, 1980.

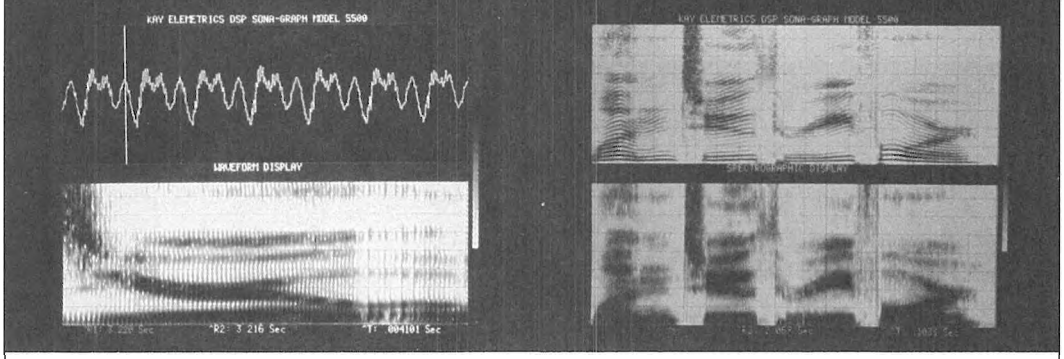
[7] D. Kewley-Port and D. Pisoni, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Am.*, vol. 73 (5), pp. 1779-1793, May 1983.

[8] S. Blumstein, "Acoustic invariance in speech production: evidence from measurments of spectral characteristics of stop consonants," *J. Acoust. Soc. Am.*, vol. 66 (4), pp. 1001-1017, Oct. 1979.

[9] J.P Cordeau, "Un système de reconnaissance- Analyse de quelques distances," Technical Repport of E.N.S.T, France, Feb. 1988.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust.. Speech and Signal Processing.*, vol. ASSP-28, pp. 357-366, 1980.

Real-time speech analysis workstation.....



....why wait?

The DSP Sona-Graph™, model 5500 is a workstation that provides state-of-the-art speech and voice analysis in a high speed environment. No waiting is required because the analysis occurs in real-time. Speech Pathologists, ENT physicians, Phoniatrists, Linguists and other speech professionals now have access to a speech workstation designed with them in mind. So.....why wait.

- ☑ **Real-Time (DC-32,000 Hz)**
- ☑ **Dual channel analysis/ display**
- ☑ **High resolution graphics**
- ☑ **Menu-driven operation for ease of use**
- ☑ **High speed computer interface**

.....And available programs keep growing for the DSP Sona-Graph.

- ☑ **LPC analysis/synthesis**
- ☑ **Voice pathology analysis (jitter, shimmer, H/N ratio)**
- ☑ **Long term spectral averaging**

For more information on using the DSP Sona-Graph in your work, call Kay's Product Specialist at (201) 227-2000 or write to the address listed below.

KAY

Kay Elemetrics Corp.
12 Maple Avenue • Pine Brook, NJ 07058
Tel: 201/227-2000 • TWX: 710/734-4347
FAX: 201-227-7760

DSP Sona-Graph™ is a trademark of Kay Elemetrics Corp.