Proceedings of The Institute of Acoustics

AUTOMATIC SPEECH RECOGNITION USING LOCAL TIMESCALE VARIABILITY INFORMATION

M J RUSSELL, R K MOORE AND M J TOMLINSON

ROYAL SIGNALS AND RADAR ESTABLISHMENT, MALVERN

INTRODUCTION

Non-linear time-normalisation, using Dynamic Programming, is now an established method for overcoming temporal variability problems in Automatic Speech Recognition. The purpose of this technique is to compute a time-registration path which defines the relationship between the timescales of two samples of speech. Much of the existing work in the field is concerned with placing restrictions on the shape of this path in order to exclude unlikely timescale distortion (for example [1],[2]). With few exceptions [3],[4], these restrictions take the form of fixed slope constraints. However, it is evident that the likelihood of timescale distortion is not fixed, but varies throughout the duration of an utterance. Hence, it is reasonable to assume that the performance of a Dynamic Time-Warping algorithm would improve if the constraints placed on the shape of the path were derived from knowledge about local timescale variability.

Recent work in this laboratory has led to the development of an algorithm for automatically acquiring such knowledge [5]. However, before attempting to incorporate timescale variability information into the Dynamic Time-Warping process, it is desirable to investigate the extent to which this information can provide a cue for classification. This paper presents such an investigation.

MEASURING LOCAL TIMESCALE VARIABILITY

The algorithm used to measure timescale variability is a modification of that derived in [5]. In the notation of [5], let $V = \{V(i):1 \leq i \leq I\}$ and $H = \{H(j):1 \leq j \leq J\}$ be samples of speech, P a set of positive simple productions and d a suitable metric. R_+ denotes the set of non-negative real numbers and $X = \{(i,j):1 \leq i \leq I,\ 1 \leq j \leq J\}$ is the (i,j)-plane. The cumulative distance CD(V,H) between V and H is calculated, by Dynamic Programming, using the forward-pass recursive equation

$$D_+(i,j) = \min_{(p,q)\in P} \{D_+(i-p,j-q) + d(V(i),H(j))\}$$

subject to initial condition $D_+(1,1) = d(V(1),H(1))$. Then $CD(V,H) = D_+(I,J)$. The local behaviour of optimal time-registration paths at a point (i,j) is described by the set $Prod_+(i,j)$, where a production (p,q) in P is included in $Prod_+(i,j)$ if and only if

$$D_+(i,j) = D_+(i-p,j-q) + d(V(i),H(j))$$

Optimal time-registration paths between V and H are obtained by back-tracking from (I,J) to (1,1) according to the productions in $Prod_+$.

Analogously, backward-pass cumulative distances are obtained from the recursive

AUTOMATIC SPEECH RECOGNITION USING LOCAL TIMESCALE VARIABILITY INFORMATION

equation

$$D_-(i,j) = \min_{(-p,-q)\in P} \{D_-(i+p,j+q) + d(V(i),H(j))\}$$

where $D_-(I,J) = d(V(I),H(J))$.

The reader should note that for each (p,q) in P and each point (i,j), the sum $D_+(i-p,j-q) + D_-(i,j)$ is precisely the accumulated distance along the best time-registration path between V and H which joins the points $(i-p,j-q)$ and (i,j) using the production (p,q). In particular, $D_+(i-p,j-q) + D_-(i,j) \geq CD(V,H)$, with equality if and only if this path is optimal.

Let $f:R_+ \to [0,1]$ be a non-increasing function such that $f(0) = 1$. Then the function $\phi_f: P \times X \to [0,1]$ defined by

$$\phi_f [(p,q),(i,j)] = f(D_+(i-p,j-q) + D_-(i,j) - CD(V,H)),$$

$((p,q)\in P,\ (i,j)\in X)$, has the following properties:

(a) $0 \leq \phi_f [(p,q),(i,j)] \leq 1$

(b) If an optimal path passes through the points $(i-p,j-q)$ and (i,j) using the production (p,q), then $\phi_f [(p,q),(i,j)] = 1$.

Hence $\phi_f [(p,q),(i,j)]$ is a measure of how well the production (p,q) explains the timescale distortion between V and H at (i,j).

The number of occurrences of the production (p,q) at the jth constant of H, $\tau^{HV}_{(p,q)}(j)$, is defined as

$$\tau^{HV}_{(p,q)}(j) = \sum_{i=1}^{I} \phi_f [(p,q),(i,j)].$$

Finally, the relative frequency of use of the production (p,q) at the jth instant of H, with respect to a set of utterances $V_1,\ldots V_N$ is given by

$$\sigma^{H\{V_1,\ldots,V_N\}}_{(p,q)} = \frac{\sum_{n=1}^{N} \tau^{HV_n}_{(p,q)}}{\sum_{n=1}^{N} \sum_{(r,s)\in P} \tau^{HV_n}_{(r,s)}}$$

For an interpretation of the functions $\sigma^{H\{V_1,\ldots,V_N\}}_{(p,q)}$ see [5].

CLASSIFICATION USING PATH-VALIDITY MEASURES

Consider the problem of classifying an unknown utterance S into one of M classes $c_1,\ldots,c_M$, where each class c_v is represented by a single reference utterance R_v $(v=1,\ldots,M)$. The standard solution to this problem, using Dynamic Time-Warping,

is to compute $CD(S,R_v)$ for each v (v=1,...,M). S is then assigned to class c_{v_0} according to a nearest-neighbour rule, ie

$$v_0 = \underset{v=1,\ldots,M}{\text{argmin}} \{CD(S,R_v)\}.$$

Now suppose that, for each class c_v a set θ_v of training utterances is available in addition to the reference utterance R_v. For each v define $\bar{\theta}_v = \bigcup_{s\neq v} \theta_s$. The functions $\sigma^{R_v\theta_v}_{(p,q)}$ and $\sigma^{R_v\bar{\theta}_v}_{(p,q)}$ ((p,q)εP), computed as in the previous section, represent the relative frequency of use of the production (p,q) for the reference utterance R_v, with respect to inclass and outclass matches respectively (see Figure 1).

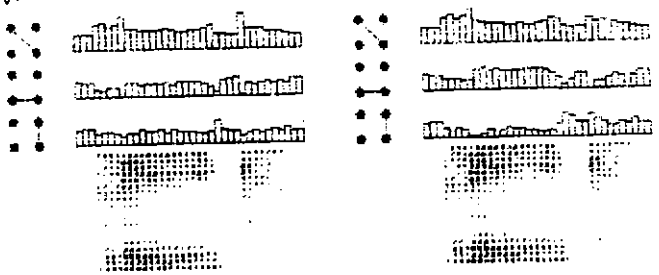

FIGURE 1 : Relative frequency of use of the productions (1,1), (0,1) and (1,0) for the utterance /li:g/ wrt 10 examples of /li:g/ (left) and /li:k/ (right).

There are a number of ways in which $\sigma^{R_v\theta_v}_{(p,q)}$ and $\sigma^{R_v\bar{\theta}_v}_{(p,q)}$ can be used to measure the extent to which a time-registration path w between S and R_v represents a likely distortion of the timescale of R_v. Such a measure will be called a path validity measure. In this paper only two path validity measures will be considered.

Fix v, and let $w = (w_1,w_2)$ be a time-registration path of length L between S and R_v (see [5]). The sequence of productions $P_2,\ldots,P_L$ which constitutes w is given by $P_n = (w_1(n)-w_1(n-1),\ w_2(n)-w_2(n-1))$, n=2,...,L.

The two path validity measures which will be considered are defined as follows:

$$\Lambda_1(w) = \frac{1}{L-1}\sum_{n=2}^{L} \sigma^{R_v\theta_v}_{P_n} \quad , \qquad \Lambda_2(w) = \prod_{n=2}^{L} \frac{\sigma^{R_v\theta_v}_{P_n}}{\sigma^{R_v\bar{\theta}_v}_{P_n}}$$

Note that Λ_1 uses only information about the use of productions within the class c_v, whereas Λ_2, based on the likelihood ratio from probability theory, requires information about the use of productions across classes.

The procedure for classifying S on the basis of a path validity measure Λ is

straightforward. For each v, the forward-pass Dynamic Time-Warping process yields an optimal time-registration path w_v between S and R_v (v=1,...,M). S is assigned to class c_{v_0} where

$$v_0 = \underset{v=1,\ldots,M}{\text{argmax}} \{\Lambda(w_v)\}.$$

RESULTS

In the following experiments, acoustic analysis was performed using a 19 channel vocoder, producing one vector every 20 ms. The metric d was Euclidean distance, P the set {(1,0),(1,1)(0,1)} and f the function defined by f(x) = max {(10-x)/10,0}.

Table 1 shows the results of an experiment to compare the performance of the above path validity measures with that of a standard Dynamic Time-Warping approach. For completeness, the standard algorithm was applied both with and without slope constraints. The task was to classify each utterance in a test set into one of two classes c_1 and c_2. For each class, a reference pattern was selected and ten training utterances (which were not in the test set) were provided for timescale variability analysis. The test set consisted of fifty utterances from each class.

The pairs (/li:g/,/li:k/),(/pɒd/,/pɒt/),(/rɑɪdə/,/rɑ:tə/) and (/kləʊz/,/kləʊs/) were chosen because, in each case, vowel duration provides a cue for discrimination. For the first three pairs, this leads to significantly lower error rates when classification is based on either of the path validity measures. By contrast, the difference between the pair (/fɑɪv/,/nɑɪn/) is due almost entirely to spectral shape and this is reflected in high error rates for classification by path validity.

word pair	cumulative distance	cumulative distance (slope const)	Λ_1	Λ_2	cumulative distance & Λ_1	cumulative distance & Λ_2
/li:g/,/li:k/	33	18	8	4	6	1
/pɒt/,/pɒd/	43	35	9	4	3	4
/rɑɪdə/,/rɑɪtə/	40	33	40	0	30	0
/kləʊz/,/kləʊs/	10	5	33	20	1	3
/fɑɪv/,/nɑɪn/	15	12	52	31	14	13

TABLE 1 : % error rates for classification by cumulative distance and path validity. (The slope constraint in column 3 corresponds to case P = 1 in [1].)

A detailed examination of the errors which occurred for the pair (/kləʊz/,/kləʊs/) revealed that those errors resulting from classification by minimum cumulative distance were different, in general, from those for classification by maximum path validity. Since this phenomenon is clearly important, a new classification rule was tested. Let R_1, R_2 be reference utterances for the classes c_1, c_2 respectively and let S be a test utterance. S is now classified as follows: If $|CD(S,R_1) - CD(S,R_2)| > K$, for some fixed constant K, then S is classified according to minimum

cumulative distance, otherwise, the cumulative distances are rejected and S is classified according to maximum path validity. K is called the cumulative distance reject threshold. The result of applying this classification rule to the pair (/kləʊz/,/kləʊs/), for varying K, is shown in Figure 2. The graph clearly indicates that a classification rule which combines spectral distance and timescale variability information can perform significantly better than one based on spectral distances alone.

The final two columns of Table 1 show the minimum errors rates obtained by applying this new rule to the remaining test data.

CONCLUSION

The preliminary experiments described above indicate that timescale variability information can be used to significantly improve recognition accuracy. Considerable research effort is now being directed towards a study of different techniques for incorporating this information into the Dynamic Time-Warping process.

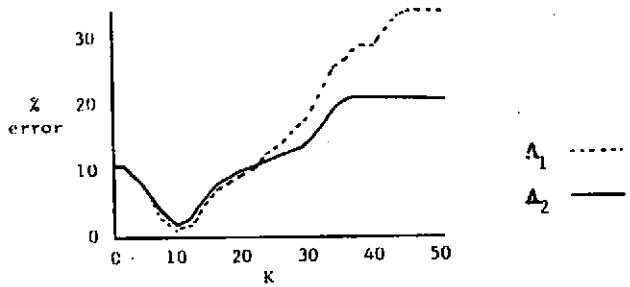

FIGURE 2 : Error rates for the 'cumulative distance reject threshold' classification rule applied to the pair (/kləʊz/,/kləʊs/).

REFERENCES

1. H SAKOE and S CHIBA 1978 IEEE ASSP 26, 43-49. Dynamic Programming Algorithm Optimisation for Spoken Word Recognition.
2. C MYERS, L R RABINER and A E ROSENBERG 1980 IEEE ASSP 28, 623-635. Performance Tradeoffs in Dynamic Time-Warping for Isolated Word Recognition.
3. R K MOORE 1981 Proc. of the Institute of Acoustics, Spring Conf., 269-272. Dynamic Programming Variations in Automatic Speech Recognition.
4. H F SILVERMAN and N R DIXON 1980 Proc. IEEE Int. Conf. ASSP, 169-171. State Constrained Dynamic Programming (SCDP) for Discrete Utterance Recognition.
5. R K MOORE, M J RUSSELL and M J TOMLINSON 1982 Proc. IEEE Int. Conf. ASSP, 1270-1272. Locally Constrained Dynamic Programming in Automatic Speech Recognition.