Proceedings of The Institute of Acoustics

AN ISOLATED WORD RECOGNITION SYSTEM WITH PROGRESSIVE ADAPTATION OF TEMPLATES

F.R. McInnes, M.A. Jack and J. Laver

Centre for Speech Technology Research, University of Edinburgh

INTRODUCTION

The use of whole-word templates (reference patterns), obtained by a training procedure from utterances of the words to be recognised, is a well-established and successful approach to automatic recognition of isolated words from a small to medium-sized vocabulary. Each unknown input word is compared with all the stored templates, and is recognised as the word whose template yields the smallest value of a "word distance" or dissimilarity measure. Each template, and each input word, is represented for this comparison by a sequence of vectors of acoustic parameters (such as bandpass filter energies or cepstral coefficients), each vector being derived from a short time segment of the speech signal [1,2].

There are various problems which arise with this word recognition procedure, because of the degree of variability that occurs among utterances of the same word, by the same speaker on different occasions or (even more) by different speakers.

Temporal variation, and DTW matching

One form of variability is in the timescale of a word. The overall duration of the word varies from one utterance to another; also the relative durations of its parts (e.g. phones or syllables) vary. To cope with this temporal variation, the comparison procedure employs the dynamic programming technique known as dynamic time warping (DTW) [1,3], which finds the optimal alignment of a given pair of input and reference patterns, together with the corresponding word distance.

The main drawback of DTW is that it is computationally expensive; the amount of computation required is directly proportional to the number of templates to be matched, and to the square of the number of vectors per word. Various modifications have been proposed to reduce the computational requirements. Among these is the application of a relatively simple preliminary comparison to eliminate templates which are very dissimilar to the input word, so that only the most likely candidates are subjected to full DTW matching [4]. It is also possible to reduce the computation for each DTW matching operation by first compressing the representation of each word to a small number of acoustic vectors: various segmentation techniques exist which can be used to accomplish this [5,6].

These ideas of segmentation to compress word representations and of elimination of unlikely words by a simple comparison can be combined, as described below, to build a multiple-stage recognition system which achieves a substantial reduction of the time required to recognise each word, with little or no loss of accuracy, relative to the basic single-stage DTW-based recogniser.

Other forms of variability, and template adaptation

The effectiveness of a template-based word recogniser depends on its having good templates for all the words in the designated vocabulary. If the vocabulary is small, and the speaker and conditions are consistent, this can be achieved by deriving a template for each word from several utterances provided by the prospective user of the system during an initial training (enrolment) session [7]. However, if

the vocabulary is large, or there are frequent changes of speaker, this requirement of training becomes burdensome and time-consuming. In these cases, an alternative is to use a speaker-independent set of templates, generated by a representative set of speakers [8]. A speaker-independent system, to achieve satisfactory performance, requires several templates per word [8]; this, however, increases the computational requirements for the template matching process. A further disadvantage of a speaker-independent recogniser is that, when a new word is added to the vocabulary, it must be spoken by a representative set of speakers to train the system, in order to maintain the desired standard of recognition accuracy.

A method of improving the performance of a suboptimally trained word recognition system, whether speaker-trained or speaker-independent, is to incorporate adaptation of the templates during the recognition session [9]. The user can start using the recogniser with a small set of speaker-independent templates, or a set of single-utterance templates generated in a short training session, and the system will improve the templates by adapting them to the recognised input, so that its accuracy increases as it is used. This adaptation will also keep track of gradual changes in the speaker's voice or the background noise or transmission conditions.

The adaptation can be supervised (conditional on feedback as to the correctness of the recognition) or unsupervised. It may be helpful, especially in the case of unsupervised adaptation, to impose some condition as to the closeness of the word match or the certainty of the recognition decision before allowing a word to be used in adaptation of the best-matching template. A further option in the case of supervised adaptation is to implement negative adaptation in instances of incorrect recognition, so that the template becomes less similar to the input word which has been misrecognised, thus making the recurrence of the same error less likely.

Various techniques for template adaptation have been proposed [9,10]. The technique considered in this paper is a fairly straightforward one, in which a weighted averaging process is applied to the existing template and the input word. This adaptation technique has been incorporated into the multiple-stage recogniser already mentioned. The remaining sections of this paper contain a description of the system, the results of some preliminary experiments into possible adaptation options and an indication of directions for intended further research.

WORD RECOGNITION SYSTEM

The overall structure of the recognition system is shown in figure 1. The subsections below describe the components of the system and the operation of the multiple-stage decision and adaptation procedures.

Data acquisition, acoustic analysis and endpoint detection

The system is implemented in software on a Masscomp MC5500 minicomputer, using a built-in analogue-to-digital convertor for data acquisition, and an AP501 array processor to perform acoustic analysis. During training, interactive recognition or test data collection, the speaker is prompted, by visual and audible signals from a terminal, to utter each word during an interval of 1.5s. The speech is low-pass filtered at 8kHz, and digitised at a 20kHz sampling rate. The beginning and end of the word are located automatically using thresholds on energy and zero-crossings in 10ms frames. (If the number of words detected in the 1.5s interval is not exactly 1, the speaker is prompted to repeat the word.) After this endpoint detection, the speech signal is subjected to preemphasis (factor 0.98), and to 8th-order LPC analysis in a 25.6ms Hamming-windowed frame every 10ms, and 8 cepstral coefficients are derived to represent each frame.

AN ISOLATED WORD RECOGNITION SYSTEM

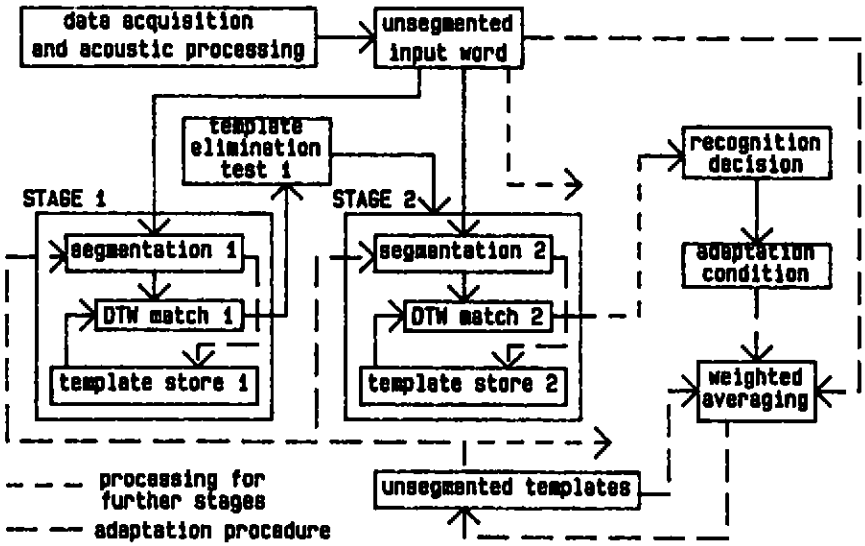
Word comparison technique

In each stage of the recognition process, the input word is segmented and compared with the (similarly segmented) templates for the words of the vocabulary under consideration. (At the first stage, all the templates are used; at later stages, some of them may have been eliminated.)

The segmentation technique involves dividing each word into a fixed number, N, of segments, and either averaging the acoustic vectors in each segment (so that the pattern after segmentation consists of N vectors) or interpolating a vector at each segment boundary (which generates N+1 vectors, including those at the beginning and end of the word). Either linear time segmentation or a form of trace segmentation [5] can be used. (Previous experiments comparing these segmentation techniques have been reported elsewhere [11].)

The segmented input word is compared with each template by DTW using Itakura's form of local path constraints and type (c) weighting [3], with the input word along the x-axis. The vector distance function in the DTW matching is the absolute value distance. (A pseudotemplate frame [12], with a constant distance to any input vector, which can be matched to any number of successive input vectors, is optionally appended before and after each template, to adjust for the possible inclusion by the endpoint detector of intervals before and after the input word.) This results in a word distance for each template matched.

FIGURE 1: STRUCTURE OF RECOGNITION SYSTEM



Multiple-stage decision procedure

The system incorporates a number of word comparison stages with different segmentation parameters. The number of stages, the details of each stage and the condition after each non-final stage for passing on templates for further comparison can easily be adjusted each time the recogniser is used. For the experiments reported here, the number of stages was fixed at 3, with segmentations resulting in 2, 10 and 30 vectors per word; the pseudotemplate frame technique was included in

the DTW at the third stage.

Appropriately segmented versions of all the templates are derived at the beginning of the recognition session and stored for use in the comparison stages. When the input word has been processed by the acoustic analysis into a sequence of vectors, this (unsegmented) word pattern is stored temporarily. In the first stage, the first segmentation is applied to the input word and it is matched by the DTW algorithm to the first segmented version of each template. The output of this stage is a set of word distances, one for each template. Let the distance obtained for template v be D_v; and let v* be the value of the template index v that minimises D_v. Then the condition for passing template v on to the next stage is that

$$D_v < t_1 D_{v^*}, \qquad (1)$$

where t_1 (>1) is the threshold for the first stage. If only one value of v (i.e. v*) satisfies (1), the input word is recognised as the word represented by template v*. In this case, the remaining stages are not required for this word. Otherwise, the second segmentation is applied to the input word, and it is compared with the second segmented version of each template whose index v satisfies (1).

The output of the second stage is, like that of the first, a set of word distances. If there is no third stage in use, the input word is now recognised as the word whose template yields the smallest word distance in the second-stage comparison. If there is a third stage, a template retention criterion similar to (1), with a different threshold t_2, is applied to the second-stage word distances, and the templates satisfying this condition are passed on to the third stage. As before, if only one template satisfies the condition, the recognition decision is made and no further input segmentation or comparison is required.

Subsequent stages, if these exist, are similar to the second stage: at each stage, the appropriate segmentation is applied to the input word, and it is compared with the similarly segmented versions of those templates not eliminated by preceding stages. At some stage a recognition decision is reached.

It is possible to include a "rejection" or "no recognition" option, in which no recognition decision is made for the current input word if at any stage the ratio of the second-best to the best word distance is less than a set threshold. The rejection threshold can take a different value at each stage.

Template adaptation

The recognition procedure, when the recogniser is being used in its primary, interactive mode, is as follows. Once an input word has been recognised, the recognised word is printed out on the terminal screen. If the verification option is in use, the user is prompted for an indication of the correctness or incorrectness of the recognition. If it is incorrect, the second-best candidate word is displayed, and again the user is asked to verify its correctness. When a recognition is acknowledged as correct, or when both the best and the second-best candidates have been dismissed as incorrect, the system prompts for the next input utterance.

There is also a simulation option (used for the experiments described below), in which verification is achieved using a table of input word identities.

Template adaptation is applied whenever a recognition decision is reached and certain conditions are satisfied. Conditions which may be imposed are the following:-

Correctness of recognition: as confirmed by the user's response, or by reference to the input word identity file.

Word distance ratio: the ratio of the second-best to the best word distance, at the stage at which the decision is reached, must exceed a threshold. (This

threshold is not specified separately for each stage of comparison; but higher thresholds at the earlier stages are in effect imposed by specifying sufficiently high thresholds for template elimination in the recognition procedure.)

The main purpose of the distance ratio condition is to prevent adaptation in cases where there is no verification of the recognition and the degree of certainty of its correctness is low.

If there is verification (i.e. the adaptation is supervised), so that the correctness condifion can be imposed, then there is also an option of negative adaptation, to make the template less like the misrecognised word; and not only the best candidate template, but also (where the best candidate is incorrect) the second-best, can be adapted, positively or negatively depending on whether it is correct.

The adaptation procedure consists of DTW alignment of the unsegmented versions of the template to be adapted and of the input word, and weighted averaging of each pair of vectors thus matched together, and interpolation of the vectors of the adapted template at integer points on a weighted-average timescale, as described in [7]. The weight on the input word is a constant, W, in the range from 0 to 1; the weight on the existing template is 1-W. (W=0 corresponds to no adaptation; W=1, to replacement of the template by the input word.) In negative adaptation, the procedure is the same, but W is negative (and so 1-W is greater than 1). When a template has been adapted, segmented versions of it are derived, replacing the previous versions, for use at all the stages of the recognition procedure.

EXPERIMENTS AND RESULTS

The adaptive recognition experiments reported here involve speaker-dependent recognition of utterances of the 10 English digits. Results have been obtained, to date, for two male speakers.

Speech data

Each set of templates, consisting of one for each digit, was formed in an interactive training session by a robust averaging procedure with DTW alignment. (The average number of utterances required per word of the vocabulary, to obtain two sufficiently similar ones, was about 3.) In these experiments two sets of templates for speaker 1 (designated R1A and R1B) and one set for speaker 2 (R2) were used.

The test data for adaptive recognition consisted of digit utterances collected in sets of 50 on separate occasions using the automatic data collection procedure mentioned above. The same sequence, containing 5 repetitions of each digit, was displayed and pronounced in each of these data collection sessions. For speaker 1, there were 10 data collection sessions over a period of nearly three weeks, providing 500 test utterances (designated T1). For speaker 2, 300 utterances (T2) were obtained in 6 sessions on successive working days. In each case, the templates were formed during the first few days of the data collection period.

All utterances, both for template formation and for testing, were recorded using a Sennheiser HME1019 headset microphone in a computer terminal room. There was a low level of continuous background noise, and there were also people working at nearby terminals during some of the sessions.

Adaptation parameters and results

The words in data set T1 were recognised using each of template sets R1A and R1B, and those in T2 were recognised using R2. The templates were adapted during the recognition process. Figures 2 and 3 show, for various adaptation

parameter values, the average word recognition accuracies obtained (over 1300 recognitions in all: 500, 500 and 300 with the respective template sets). Figure 2 shows results for adaptation with verification, with no distance ratio condition, with a number of combinations of positive and negative adaptation weight values. Figure 3 shows the performance using adaptation without verification, with and without a distance ratio condition. In the cases with verification, the second-best candidate template was also adapted when the first candidate was incorrect. The performance with no adaptation is shown as the first point marked "0" in each figure.

FIGURE 2:
RESULTS WITH SUPERVISED ADAPTATION

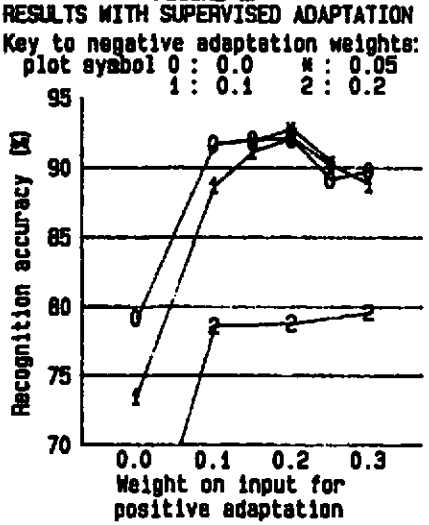
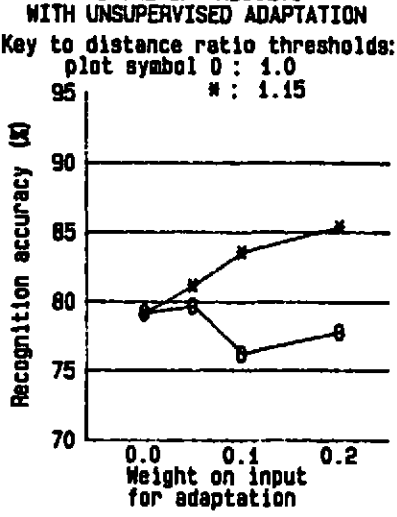


FIGURE 3: RESULTS WITH UNSUPERVISED ADAPTATION



The best recognition performance was obtained using supervised adaptation, with weights of 0.2 and -0.05 on the input in positive and negative adaptation. The best result for unsupervised adaptation was obtained when the adaptation was conditional on a word distance ratio exceeding 1.15 and the input weight was 0.2. (These values may not be optimal, as only two distance ratio thresholds and three weights have been tested for unsupervised adaptation.) The rates of correct recognition were 79.2% without adaptation (69.2%, 91.0% and 76.3% for the individual combinations of test data and templates); 92.8% (94.2%, 95.0%, 87.0%) with the optimal supervised adaptation; and 85.4% (85.2%, 92.4%, 74.3%) with the best unsupervised adaptation. The improvement is greater for R1A than for R2A: before adaptation the performance of R1A was considerably poorer, but with adaptation the two template sets yielded similar results. The poor results for speaker 2, even with adaptation, suggest that many of the errors for this speaker were due to deficiencies in the test data rather than in the templates.

DISCUSSION

The results obtained thus far indicate the usefulness of the template adaptation technique in enhancing speaker-dependent isolated word recognition performance. In particular, the adaptation procedure with verification can improve the

performance of a poor set of templates (such as R1A) to an apparently near-optimal level. More detailed examination of the results indicates that most of the improvement in the templates has occurred after about 50 input words. (This suggests that about 5 utterances of each word of the vocabulary are required for effective adaptation; but further experiments will be necessary to establish a more accurate estimate.)

The negative adaptation for misrecognised input appears to be of some benefit, but only if the negative weight is kept small and it is used in conjunction with the positive adaptation. It might be helpful to impose some word distance condition on negative adaptation, or to apply it only where the second-best candidate was correct: this could prevent adaptation away from noisy or badly detecteed input.

Even without feedback for verification, template adaptation can still improve the system's performance - though in this case it is preferable to have a threshold imposed on the ratio of the best two word distances, to prevent adaptation in cases of uncertainty. The choice of the distance ratio threshold is significant: if it is set too low, there is a risk that a template will be adapted repeatedly to utterances of the wrong word, resulting in severely degraded recognition performance on the misrecognised word and on the word that the template is intended to represent. More extensive experiments will be required to show whether it is possible to prevent this instability from arising over long sequences of input words. (If this cannot be guaranteed, a retraining procedure will have to be provided: see below.)

Further research is planned to extend the above results to more repetitions of the same words; to other vocabularies; to a larger number of speakers; to isolated word recognition using speaker-independent initial templates; and to connected word recognition, with initial templates derived from isolated utterances or a limited set of embedded utterances. There are also various options using multiple templates which could be explored: for instance, in a connected word recognition system, the adaptation procedure might be employed to generate from each word's initial template a set of adapted templates corresponding to contextual variations.

In applications of speech recognition, an important field of investigation is the interaction between the user and the system. The interactive recognition mode of the system that has been developed will allow experiments to be carried out in which the user can adapt to the recogniser as well as vice versa. The interactive mode allows more flexibility in the operation of the system, as the user can repeat words which are wrongly recognised, and, in the event of repeated failure to recognise a particular word, can abandon an existing template and generate a new one by providing one or more fresh training utterances of the word. (In practice, a user of a word recognition system is unlikely to tolerate very poor performance on particular words of the vocabulary - especially if each word has to be repeated until it is recognised correctly. So it is preferable to have a retraining procedure available for use as required during the recognition session.) The assessment of a system's performance becomes more complex as the degree of interaction between system and user increases; but this interaction is such an important feature of any application of a speech recogniser as to merit investigation despite this difficulty.

SUMMARY

An implementation of an isolated word recognition system incorporating a multiple-stage decision procedure and template adaptation has been described. Preliminary results of experiments with this system indicate that the template adaptation procedure can greatly improve recognition accuracy, especially where the initial set of templates gives poor performance. There is scope for further

investigation of a number of aspects of the adaptive recognition process, and for application of the adaptation technique to speaker-independent and multiple-template systems and to connected word recognition.

ACKNOWLEDGEMENT

The work reported in this paper was made possible by support from the Science and Engineering Research Council.

REFERENCES

[1] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-23, 67-72 (1975).

[2] G.M. White and R.B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-24, 183-188 (1976).

[3] C.S. Myers, L.R. Rabiner and A.E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-28, 623-635 (1980).

[4] T. Kaneko and N.R. Dixon, "A Hierarchical Decision Approach to Large-Vocabulary Discrete Utterance Recognition", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-31, 1061-1066 (1983).

[5] M.H. Kuhn and H.T. Tomaschewski, "Improvements in Isolated Word Recognition", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-31, 157-167 (1983).

[6] R. Pieraccini and R. Billi, "Experimental Comparison Among Data Compression Techniques in Isolated Word Recognition", Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., April 1983, 1025-1028.

[7] R. Zelinski and F. Class, "A Learning Procedure for Speaker-Dependent Word Recognition Systems Based on Sequential Processing of Input Tokens", Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., April 1983, 1053-1056.

[8] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg and J.G. Wilpon, "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-27, 336-349 (1979).

[9] R.I Damper and S.L MacDonald, "Template Adaptation in Speech Recognition", Proc. IOA, vol. 6, 293-299 (1984).

[10] Y. Niimi and Y. Kobayashi, "Synthesis of Speaker-Adaptive Word Templates by Concatenation of the Monosyllabic Sounds", Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., April 1986, 2651-2654.

[11] F.R. McInnes, M.A. Jack and J. Laver, "Comparative study of time segmentation and segment representation techniques in a DTW-based word recogniser", IEE Conf. Pub. 258 (Speech Input/Output; Techniques and Applications), 21-26 (1986).

[12] J.S. Bridle, M.D. Brown and R.M. Chamberlain, "An Algorithm for Connected Word Recognition", Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., May 1982, 899-902.