Proceedings of The Institute of Acoustics

TEMPLATE ADAPTATION IN SPEECH RECOGNITION

Robert I. Damper and Sara L. MacDonald

Department of Electronics and Information Engineering, University of Southampton.

INTRODUCTION

Useful application of speech-recognition devices within systems demands that recognition accuracy must be high and training by the speaker must not be unnecessarily complicated or tedious. Indeed, to be practical, any man-machine system must allow the user to achieve his goals with 100% success. However, no automatic speech recogniser can of itself achieve the required performance: not even humans can recognise speech with perfect accuracy and, unavoidably, speakers will occasionally mis-speak. Therefore, adequate error detection and correction procedures must be incorporated to convert an imperfect device into a 100% accurate system: the provision of appropriate feedback to the user is an effective way of allowing errors to be detected. Since error correction is essentially unproductive, however, users should not be required to invoke these procedures too often; a consideration which leads to the notion of an acceptable error rate.

One technique is often suggested for keeping errors within acceptable limits - the use of adaptive recognition. For instance, Underwood [1] states:

> "In designing a system that will interact with a man by means of speech, one cannot regard the user as having characteristics that do not change with time. For instance, fatigue may affect a user's voice pattern. ... Thus, a good speech recogniser needs to be responsive to the kind of changes a person is likely to make to his voice when speaking."

One way to make the recogniser responsive to such changes is to use recent input utterances to update the stored templates, thereby improving recognition accuracy. This adaptive approach to recognition offers other possible benefits in addition to reduced error rates - such as simplifying prior training and giving a degree of speaker independence.

In this paper, we make some general remarks on adaptive speech recognition before discussing a particular template adaptation scheme which we have implemented. We outline finally the many possibilities for useful future work. Before considering adaptation in the specific context of speech recognition, however, it is worthwhile making a few comments about adaptive man-machine interfaces in general.

ADAPTIVE MAN-MACHINE INTERFACES

In his seminal work, Wiener [2] considered the mechanisms by which biological systems adapt to their environment - so-called ontogenetic learning. Wiener was also interested in the possibility of building machines which could mimic the adaptive, or learning, behaviour of living systems. He concluded that:

> "In general, a learning machine operates by non-linear feedback."

The importance of the notion that feedback processes lie behind adaptive behaviour, in both man and machine, is that the need for stability of the updating process is emphasised. We will return to this central issue below.

The situation is considerably complicated in the case of man-machine interaction when both the man and the machine are capable of adaptive behaviour. Now the non-linear feedback necessary for adaptation may or may not be internal to the system (man or machine) itself: it could in addition be a global feature of the interface. We have already mentioned in the Introduction how this latter sort of feedback, in the form of information relayed from recognition device to speaker, is essential to good man-machine system design, even with a non-adaptive recogniser.

Edmonds [3] has considered this general case in detail, again stressing the importance of feedback (and stability) by stating:

> "If a system is to be adaptive, it must contain a mechanism for providing negative feedback [our emphasis]. In looking at the adaptive design of man-computer interfaces we need, therefore, to consider evaluation in order to provide the feedback."

Edmonds classifies the various types of adaptive interface according to the source of the feedback. In most of the cases considered, the source is a human (e.g. computer specialist, any user). In one case, however - called "self-adaptive" (see also Innocent [4]) - the interface changes automatically in response to its experience with users. Thus, at least part of the feedback is internal to the computer or its input device. It is this sort of system which is our principal concern here.

Before leaving the general topic of adaptive interfaces, one particular issue should be mentioned. Again, quoting from Edmonds [3]:

> "In all forms of adaptive interface ... a particularly difficult problem is that the human and the computer both try to adapt to one another. Such situations require very careful handling if problems are not to arise, and it would seem that considerable work remains to be done in this area."

We will return to this issue below but for the moment we note the implication that stable adaptation requires both man and machine to have access to adequate models of the interaction process.

ADAPTIVE SPEECH RECOGNITION - GENERAL COMMENTS

Adaptation is frequently suggested as a means of reducing recogniser error rates: however, little work has been reported on the topic. This is undoubtedly due to the extreme complexity of any recognition system in which both man and machine display adaptive behaviour. An understanding of such systems requires detailed knowledge of the psychology of human-computer interaction by speech, and the theory of non-linear control systems. In the absence of adequate understanding, empirical studies aimed at answering quite basic questions about the utility of adaptive recognition are valuable.

Recogniser Training

We have, so far, considered "adaptive" as synonomous with "learning". Of course, a sort of learning process - the training phase - is a necessary precursor of speaker-dependent recognition and, indeed, some writers actually refer to speaker-dependent systems as "adaptive" (e.g. Martin and Welch [5].) Although this is arguably a somewhat naive interpretation of the description, recogniser training is important in the context of adaptive systems.

TEMPLATE ADAPTATION IN SPEECH RECOGNITION

The current generation of speech recognisers are essentially statistical pattern-recognition devices, and the training process can be viewed as the system "learning" the statistics of the command utterances. The simplest case is when there is only a single training pass through the vocabulary. This simple scheme cannot be expected to work particularly well, because the high inherent variability of speech means no one training sample is ever likely to be representative of all possible occurrences of a particular utterance. In attempting to cope with this variability, it has become fairly common practice to use multiple training passes through the vocabulary, averaging the resulting templates. There is, however, some evidence from applications studies that recognition accuracy is not significantly improved as a result, in spite of the more protracted training. (See for example references [6] and [7]). This disappointing experience is no doubt due to the difficulty of ensuring that the averaged templates are indeed truly representative of the respective utterances i.e. the difficulty of ensuring stability.

An alternative, but extremely simple-minded, approach is to store multiple templates for each utterance; one set of templates for each training pass through the vocabulary. This works reasonably well in practice (as we would expect, knowing the high degree of variability of the input utterances), but the penalty of increased training time persists. In addition, template storage requirements are increased and there are more templates to search and compare to the input utterance in order to determine the best match.

With the above training procedures, therefore, it seems certain that the recogniser does not learn the statistics of the input utterances in anything like the best way. In particular, it is worthwhile considering moments of the distribution of order higher than the mean - perhaps by including variance measures as components of the feature vectors. Again, training has to be sufficiently comprehensive to allow stable estimates of the statistical moments to be made.

Distance Metrics

It is certainly possible to improve recognition performance by using in the classification process a distance metric which takes account of higher-order statistical moments. One such metric is that due to Mahalanobis [8], which incorporates a covariance-like measure. (The populations are assumed to be Gaussian.) The Mahalanobis metric has been used with success by Jesorsky for speaker identification [9], and by ourselves for isolated-word recognition [10]. Once more, these improvements are obtained at the expense of much longer, more tedious pre-training.

Integrated Recognition and Training

The benefits of basing recognition on statistical knowledge can be retained without the penalty of unacceptably protracted pre-training. Traditionally, training has been viewed as an "enrolment" process to be completed before the recognition phase can commence. However, it is somewhat incongruous to place such emphasis on collecting utterance statistics in prior training when, once recognition commences, there are any number of utterances available to the recogniser. The idea is starting to emerge that training should continue throughout use i.e. the system should be adaptive (e.g [6]). Elsewhere [11], we have argued that:

TEMPLATE ADAPTATION IN SPEECH RECOGNITION

> "... training and adaptation ... can in some senses be viewed as similar processes. Adaptation is, in effect, a training procedure which is integrated with normal operation."

However, the Mahalanobis metric appears unsuitable for such adaptive recognition (quite apart from the Gaussian assumption) for two reasons. First, updating of the covariance metric on an utterance-by-utterance basis is non-trivial. Second, the metric implicitly assumes the statistics to be stationary (time-invariant); a very questionable assumption in the case of speech. More realistically, we expect the statistical moments to drift, and it is these variations which the adaptive system should follow.

PREVIOUS WORK

As stated above, there is little reported work on adaptive speech recognition. One exception is the paper by Green et al [6], which describes significant performance improvements obtained by template adaptation. This work is interesting for the attention paid to the problem of stability. For instance, the authors state:

> "... a successful adaptive system requires a reliable error signal, to measure the discrepancy between the target and the observed output; a means to control the output to reduce the error signal; and a guarantee of stability ..."

Various ways of deriving the error signal were considered. Initially, to prove the value of adaptation, a (100% correct) error signal was supplied by the experimenter, who informed the system whether the recogniser's word identification had been correct or not. If correct, the input was used to update the respective template. Results showed that very significant performance improvements could be so obtained.

To be practical, however, some measure of "self-adaptation" is needed; with the system generating its own error signal. Green et al at first attempted to do this by only updating templates when the word identified had a low associated distance i.e. it was below a tightly-set "rejection limit". This scheme was, however, unstable - performance was poorer than for the non-adaptive case - since the probabilistic recogniser was bound to identify some utterances incorrectly but with a high goodness-of-fit level. Their most successful arrangement was rather complex, and used four separate templates for each word in the vocabulary. The total vocabulary was divided into two equal parts - each containing two templates for every word. For updating to take place, the input had to be recognised with a high degree of confidence as one version of the same word in both sub-vocabularies. Updating was effected by averaging the input utterance with a different version of the same word in one of the two sub-vocabularies, and was alternated between sub-vocabularies. With this complicated scheme, the deleterious effect of an incorrect update was contained. In the words of the authors:

> "... the effect is merely to corrupt one of the four templates in one of the two vocabularies ... that template will no longer resemble any real utterance. By putting that template out of commission for a while, further mis-recognitions were not made any more likely, while further correct identifications would gradually remove the corruption from the tainted entry."

Clearly, the disadvantages of this arrangement are the complex updating

procedure and the need to store four templates per command. These disadvantages should not be overstated, however. Quite a small set of commands is usually sufficient for quite elaborate voice-control applications, and users may not be able to remember a large command-set.

Green and his co-workers do not consider the problem, referred to above, of speaker and recogniser attempting to adapt to one another. Indeed, we have found no reference in the literature to this problem in the specific context of speech recognition. Underwood [1], however, mentions work done in his laboratory with non-adaptive recognisers which is relevant. Speakers were observed to see what they did when misrecognised. Underwood states:

> "Our preliminary findings indicate that the only stratagem adopted by speakers is to speak more loudly. If the design of the machine is such that changes in amplitude do not affect its performance, then such a stratagem will not improve the recognition performance for that user with that machine ..."

Evidently, the (presumably naive) speakers had a very inadequate conceptual model of the automatic recognition process. It seems clear that the feedback given should assist the speaker to form an appropriate conceptual model of the interaction, and this will be additionally important in the case of adaptive recognition.

CLUSTER ANALYSIS AND ADAPTIVE RECOGNITION

In previous work [10], we have examined the use of cluster analysis in isolated-word recognition. The basic notion was to cluster the input utterances into similar sets and assign a template to each. (Of course, the simple strategy described earlier of using multiple templates, one for each training pass through the vocabulary, does not of itself guarantee that each template is in any way representative.) Each cluster formed was the set of training samples for which the distance between the cluster centre and the samples was less than some "distance limit", L. In spite of the superior performance achieved, we felt this approach to be unsatisfactory for routine use, because of the necessity to collect multiple training samples and the inconvenience this would cause to the user.

The work, however, suggested that the principle of clustering input utterances could be extended - with the aim of integrating the train and recognise modes of operation in an adaptive system - so as to improve performance and simplify prior training. The approach, simplified as much as possible to ease microprocessor implementation, is:

- train the recogniser with a single pass through the vocabulary. In most cases, this sample will be adequately representative of the largest cluster. If not, this should become apparent in later use, when any words causing problems will have to be retrained.
- recognise input utterances according to a suitable classification rule. In our system, the simple Chebychev distance metric was used for classification because of its computational simplicity.
- update the template of the recognised word only if the input is "similar" in some sense to the template. The similarity criterion must be sufficiently strict to assure stability of the templates.

For a number of reasons, updating was by simple averaging. Not only is

averaging easily implemented, but the weights of the feature vectors forming the templates have the useful property of reflecting temporal order. The updating rule was:

if ((D(X,T) < L) and (X verified as correctly recognised))
then update (X,T)
else continue

Here, X is the input feature vector, T is the template of the word recognised, and D signifies the Euclidean distance. This metric was used because of its simple geometric interpretation as "distance" in the feature space. This rule constitutes the non-linear feedback by which adaptation is effected. The way that X was, in fact, verified as correctly recognised was by direct confirmation by the speaker, who responded "yes" or "no" to the recogniser's selection. (This has been called "secondary feedback" by Knight and Peckham [12].) Thus, our system is only partly self-adaptive, since the feedback is not entirely internally generated.

Experiments with this system showed that significantly lowered error rates could be achieved as a result of adaptation, although performance was sensitive to the value, L, of the distance limit. A single, global L value was used, although we believe that better performance would result from use of different L-values for each word. Plotting error rate against L revealed a distinct minimum at the optimal L value. If the updating criterion was too strict (L too small), however, performance tended towards that of the non-adaptive system. If L was too large, the templates degenerated with time and adaptation was unstable. Subsequent work showed that in the purely self-adaptive case (primary feedback only, without a confirmation phase), stable operation can be achieved if the limit, L, is set very strictly.

Our results have been obtained with only one speaker, one of the authors (SLM), and so we cannot be sure that they are generally valid. In particular, it seems certain that the optimum limit value will vary with vocabulary and speaker. We do believe, however, that our general approach is sound, and performance improvements can certainly be expected with adaptive recognition.

Problems of divergent adaptive behaviour between speaker and recogniser were not encountered: on the contrary, the adaptive interface was pleasant to use and the improved performance was subjectively noticeable. This may be because the speaker, as a system designer, was in a position to form a rather sophisticated conceptual model of the interface.

FURTHER WORK

The topic of adaptive recognition has received little serious attention, and a great deal of work remains to be done. At the simplest level, it is necessary to establish the generality of our findings by conducting experiments with a larger number of speakers.

Theoretical (rather than empirical) treatment of stability criteria is urgently required. This is particularly so if, as seems likely in view of the non-stationarity of speech statistics, parameters in the feedback equation

(updating rule) need to be time variant for optimal performance.

Our work has concentrated on isolated-word recognition using linear time-normalisation. Recently, however, we have been using non-linear normalisation (dynamic programming) to make recognition less sensitive to variability in end-point detection. We are at present considering ways of extending our adaptive methodology to allow its use with dynamic-programming classification.

ACKNOWLEDGEMENTS

This work was supported by Grant GR/C/04574, Voice Input Aids for the Physically Disabled, from the Science and Engineering Research Council.

Certain of the ideas presented here benefitted from a discussion with Mr. J.S. Bridle of the Joint Speech Research Unit, Cheltenham.

REFERENCES

[1] Underwood, M.J. (1980), "What the engineers would like to know from the psychologists", in Spoken Language Generation and Understanding, J.C. Simon (ed.), D. Reidel, Dordrecht, 69-75.

[2] Wiener, N. (1961), in Cybernetics: or Control and Communication in the Animal and the Machine, MIT Press, Cambridge, Mass. (2nd Edition).

[3] Edmonds, E.A. (1981), "Adaptive man-computer interfaces", in Computing Skills and the User Interface, M.J. Coombs and J.L. Alty (eds.), Academic Press, London, 389-426.

[4] Innocent, P.R. (1982), "Towards self-adaptive interface systems", Int. J. Man-Machine Studies, 16, 287-299.

[5] Martin, T.B. and Welch, J.R. (1980), "Practical speech recognisers and some performance effectiveness parameters", in Trends in Speech Recognition, W.A. Lea (ed.), Prentice-Hall, Englewood Cliffs, New Jersey, 24-38.

[6] Green, T.R.G., Payne, S.J., Morrison, D.L. and Shaw, A.C. (1982), "Friendly interfacing to simple speech recognisers", Behaviour and Information Technology, 2, 23-38.

[7] Damper, R.I., Lambourne, A. D., and Guy, D.P. (1984), "Speech input as an adjunct to keyboard entry in television subtitling", in Interact '84: Human-Computer Interaction, B. Shackel (ed.), North-Holland, Amsterdam, in press.

[8] Mahalanobis, P.C. (1936), "On the generalised distance in statistics", Proceedings of the National Institute for Science, Calcutta, 12, 49-55.

[9] Jesorsky, P. (1978), "Principles of automatic speaker recognition", in Speech Communication with Computers, L. Bolc (ed.), Macmillan, London.

[10] Damper, R.I. and MacDonald, S.L. (1984), "Statistical clustering procedures applied to low-cost speech recognition", J. Bio-Med. Engng., in press.

[11] Damper, R.I. (1984), "Voice-input aids for the physically disabled", Int. J. Man-Machine Studies, in press.

[12] Knight, J.A. and Peckham, J.B. (1984), in A Generic Model for the Assessment of Speech Input Applications, Report from Logica UK Ltd.