EVALUATION OF SPEECH RECOGNITION BY SYNTHESIS

Laurie Moye

Marconi Speech and Information Systems, Portsmouth
Currently on Industrial Assignment to Speech Research Unit,
DRA Malvern, Malvern, Worcs WR14 3PS, UK

1 INTRODUCTION

Hidden markov models (HMMs) form the basis of most successful speech recognisers today. They succeed despite modelling the speech signal very badly as a sequence of piecewise-linear spectral segments. They succeed because they use dynamic programming (d.p.) to make a single decision to classify each frame of input data, and because the statistical basis of the model enables it to be optimally trained by efficient techniques.

Recognition by synthesis (RbS) uses the same single d.p. decision with a far more accurate model in which the transitions between speech sounds and between words can be properly represented.

The disadvantage of the synthesis model is that it does not provide any theoretical basis for efficient training. The purpose of the present work is to investigate the problems of training an RbS recogniser, with the ultimate aim of showing that it can give a better performance than recognisers with simple stochastic models.

The aim of this paper is to show that the simplified version of the recogniser being used is adequate to investigate some of the basic problems. Because it is the first publication to describe this work, it has to include a description of the recogniser which has to be too short to be adequate. A report describing the recogniser and the experiments in greater detail is in preparation.

The recogniser used in this work has been developed from the one produced by the STL component of Alvey Project MMI/069. The work is the Marconi Speech and Information Systems component of IED project number 3/1/1057, Speech Recognition Techniques, which also involves Cambridge Algorithmica, the Defense Research Agency Speech Research Unit and Parsys Limited.

2 RECOGNITION BY SYNTHESIS

Bridle[1] and Russell et al.[2] have shown that, if the matching penalties of a d.p. template matching recogniser are considered equivalent to the log probabilities of an HMM recogniser, the two recognisers can be considered equivalent. An RbS recogniser can be considered as a template recogniser in which the values to be matched to the input data are not looked up in a table of model or template data, but generated, as required, by the synthesis system. The RbS recogniser uses d.p. with partial traceback in the same way as a template or HMM recogniser, but the on-line generation of the reference data requires a very different implementation.

It has been shown[3] that improved results can be obtained from an HMM recogniser by transforming the input speech spectral data by linear discriminant analysis so that those components of the input vector with the least discrimination ability can be discarded. The data from a speech synthesis system, in the form of frames of formant and other data for the control of a terminal (vocal tract) synthesiser, already provides a very compact set of data. It should be well suited to the discrimination of speech sounds, so this data is matched directly to the (suitably analysed) input data.

Unfortunately, unlike the parameters of an HMM, the real speech data cannot be considered to have a gaussian distribution about the values produced by the synthesis model. For example, when a formant ɹn the model has zero amplitude, its frequency, although defined, only represents the frequency of the real formant in the transitions into and out of that segment. A matching algorithm must be developed pragmatically to handle such relationships.

The input speech analysis has to provide frames of data to be matched to the formants in the reference data. To use formant extraction in the analysis would be to make a highly error-prone decision outside the single d.p. decision of the recogniser. Instead, the analysis provides the frequencies and amplitudes of all peaks in the input signal and these are matched separately to the reference formants of each different segment hypothesis within the d.p. process in the manner described by Hunt[4, 5].

2 STRUCTURE OF THE RECOGNISER

The recogniser is based on the JSRU synthesis-by-rule system. This provides a complete hierarchy of components operating at every level between unrestricted English and the speech signal. It was also familiar to us, and available to the project in source-code form.

In a fully developed RbS recogniser, the synthesis model would be provided with a dictionary and grammar defining valid sequences of words in orthographic form, and it would generate, on demand, the pronunciation, stress, segment durations and finally the sequence of frames for each hypothesis being actively evaluated by the d.p. algorithm. For this initial implementation, only the lowest level of the synthesis, the Holmes-Mattingley-Shearme (HMS) rules for interpolation of segments into frames, are incorporated into the recogniser. The permissible sequences of segments are generated off-line for a particular vocabulary and syntax and incorporated into a network grammar.

A full implementation of the HMS rules for use in a recogniser would accommodate variation in speaking rate by generating each segment with a variety of (appropriately penalised) durations. To limit the complexity of implementing the HMS rules to generate frames on demand for many hypotheses simultaneously, this recogniser generates each frame with only one (context dependent) duration and the frames are time-warped as they are matched to the input data.

3 TRAINING

The recogniser is trained by altering the talker-specific segment parameters used by the HMS rules. At present, it is only the target values in this synthesis table that are changed. The durations and percentages which control the transitions will not be adapted until more experience has been gained adapting the targets.

A new synthesis table is obtained by the same technique that was used to train the JSRU synthesis system. The training data is aligned to the reference data by forced recognition using a grammar specific to each training utterance. At present, the training grammars are generated by hand from the segment-level output of the JSRU synthesis system. The output of the recogniser is labeled time-aligned natural speech. It contains frames of input data (but with the formants now extracted) with additional lines identifying the segments to which it has been matched.

The HMS rules express each parameter of a frame as a weighted sum of target values for the segment in question and its left and right neighbours. For any given parameter, the total squared error between the real value and the synthetic value, for all the training data, can be expressed as a weighted sum of the target values of all the segments (that occur in the training data). Minimising this error by equating the partial derivatives of the total squared error (with respect to the target values) to zero gives a set of linear simultaneous equations which are solved to find the optimum targets for this training data. Segments that do not occur in the training data retain their original values.

An experiment consists of a series of iterations each requiring alignment of the training data, adaptation of the synthesis table, recognition using this new table and scoring of the results. The training data consists of many short files with one utterance each, the only annotation used being at the utterance level. Experiments are performed with, typically, six iterations, producing many hundreds of files in the process. A fully automatic method is used to run these experiments which requires only that the files containing the parameters for each stage of the process are put into a top level directory. The experiment can be run for the required number of iterations (or continued for more iterations) with a single command, the necessary file names and directories being automatically generated.

Scoring of the recognition results is done with a d.p. scoring program with no alignment penalties, and penalising mismatches, insertions and deletions equally.

All the experiments so far have used one set of 50 digit triples from the English data of Eurom0. All "evaluation" has been done on the training set so that the other set of 50 triples from this speaker is available to be used when the performance justifies it. The other English speakers will be used later to evaluate the procedure and parameters developed on this first speaker.

4 EXPERIMENTS

Experiments have been performed with the intention of optimising some of the matching parameters in the recogniser. The results showed that training of the synthesis table over about six iterations would produce improvements in performance, but the results were very poor and not particularly sensitive to the changes in parameters.

To investigate the reasons for this behaviour, scatter plots were made comparing the real and synthetic data for every one of the 9000 or so frames in the training set for the particular alignment of segments and input peaks produced in a given iteration. Normally the first iteration proved most informative because it relates the familiar values of the initial (standard) table to the input data.

It was concluded that some changes to the structure of the matching and pruning algorithms are needed before the parameters can be effectively adjusted. The details of some of the experiments are described below; experiments are identified by the serial numbers given to them by the automatic system.

4.1 Expt006 — Pruning Threshold

Pruning is controlled by a threshold; any hypotheses that score worse than the best hypothesis by more than the threshold are pruned. If the pruning threshold is too large, the recogniser runs too slowly to be practicable (even in simulation). If the pruning threshold is too small, part of what should ultimately become the best path may be pruned, thus damaging the performance.

The alignment is done with files of one digit triple each and each training grammar is a single string of segments, so a very high pruning threshold can be used, such that hardly any paths are pruned, without the recogniser running too slowly. For the evaluation run of each iteration, on the other hand, the pruning threshold has to be set as low as possible to run in a reasonable time, but high enough not to damage the recognition performance.

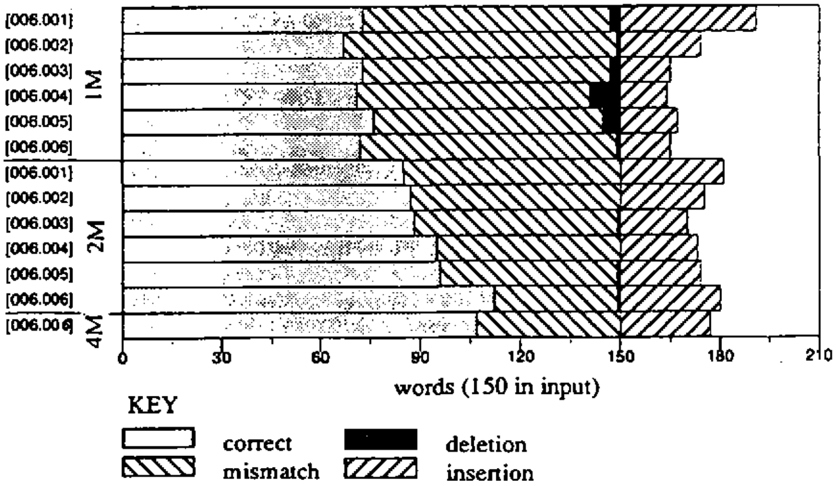

Figure 1: Results of experiments 006, the evaluation is repeated with different pruning thresholds: 1M, 2M and (for the sixth iteration only) 4M.

Expt006 was run for six iterations with evaluation pruning of 10^6 (1M) and then the evaluation was repeated with pruning of 2M and 4M. The evaluation with 4M pruning ran extremely slowly (133 Hr for a single iteration) and produced a result no better than with 2M pruning, so it was stopped. The results are summarised in fig.1. Clearly, a pruning threshold of 2M is needed to reveal the improvement brought about by the adaptation, so subsequent experiments were done with a pruning threshold of 2M for evaluation.

The inability to complete the experiment for values above 2M prevents a proper investigation of the effect of pruning threshold. This raises questions about the way pruning is done which are considered further in connection with expt009.

4.2 Expt007/008 — Energy Threshold

In the distance measure, the score due to the formants is weighted according to the energy of the input frame. Below a threshold, the formant score is progressively replaced by a score dependent only on overall amplitude differences. This threshold was set at zero in expt006 to disable this feature. Expt007 and expt00S were done with values of 20dB and 40 dB giving the results summarized in fig.2 (all using 2M pruning for evaluation). These show that, although the first iteration results are improved by the higher energy threshold, the results after 6 iterations are not significantly different.

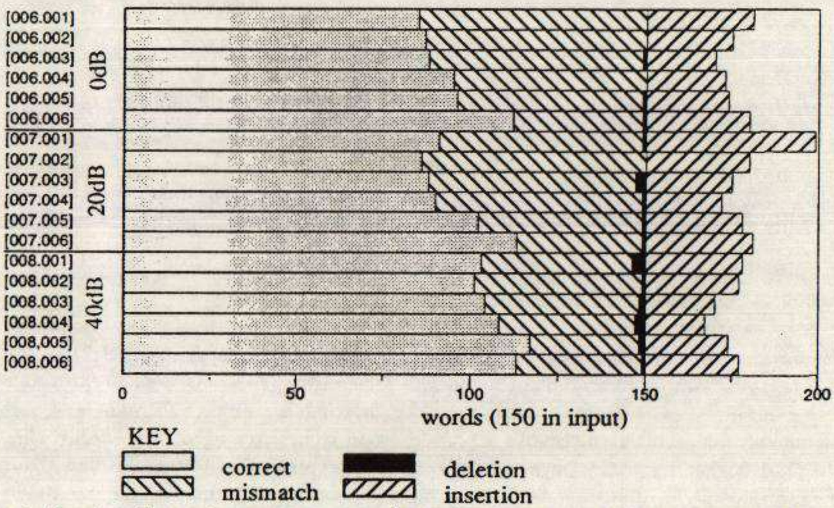

Figure 2: Results of experiments 006,007 and 008. Evaluation pruning threshold: 2M throughout. Energy threshold: 0dB for ˙006], 20dB for [007] and 40dB for [008].

Scatter plots of the first iteration alignment for each experiment showed that varying the threshold made a barely discernible difference to the results. The scatter plot for the F1 in expt006 is shown in fig.3. One would expect the points to be concentrated either side (horizontally) of a 45° line through the origin, but there are many points where a synthetic F1 value is being matched to a

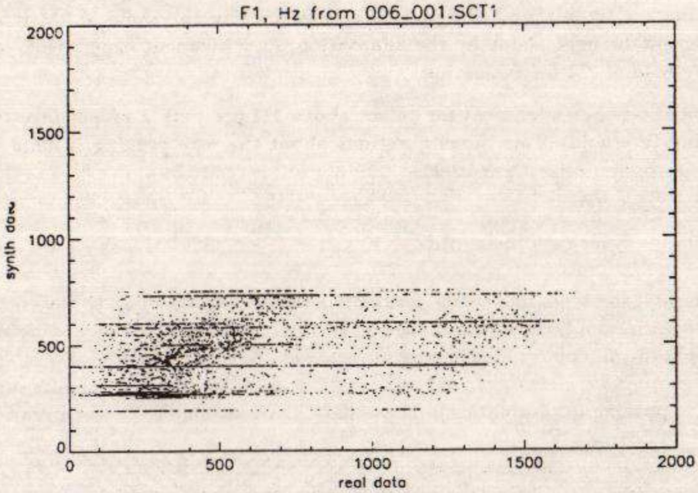

Figure 3: Scatter plot of first formant frequency for the first iteration of expt006.

peak in the input data at far too high a frequency. These points appear to belong to frames where there is no F1 peak in the input data; until the formant matching algorithm has been modified to correct this fault, it is not possible properly to adjust the energy threshold.

4.3 Expt009 — Deletion Penalty

Hunt's rules for formant matching[5] allow peaks in the input or formants in the reference data to be deleted if there is nothing to match them to. The previous experiments used fixed production penalties (the simplest initial option), a high one for deleting a reference formant and zero for deleting an input peak. This resulted, for F1 at least, in the matching of low energy reference formants to much higher frequency peaks in the real data (typically in formants). The algorithm was modified so that the fixed penalty for both input and reference deletion was augmented by the square of amplitude (measured above a threshold).

The experiment (expt009) ran extremely slowly. After three days, the experiment was stopped because it was still on the second iteration. The evaluation showed that although the words correct score was worse than expt006 (79 against 85), there were fewer insertions (22 against 31). In fact, percentage accuracy (% correct - % inserted) is higher than for expt006.

The scatter plot for F1 from the first iteration of expt009 is shown in fig.4. Compared with that for expt006, it is clear that the erroneous very high F1 values have been largely eliminated. Completion of this experiment might show much higher scores by the sixth iteration, but this cannot easily be

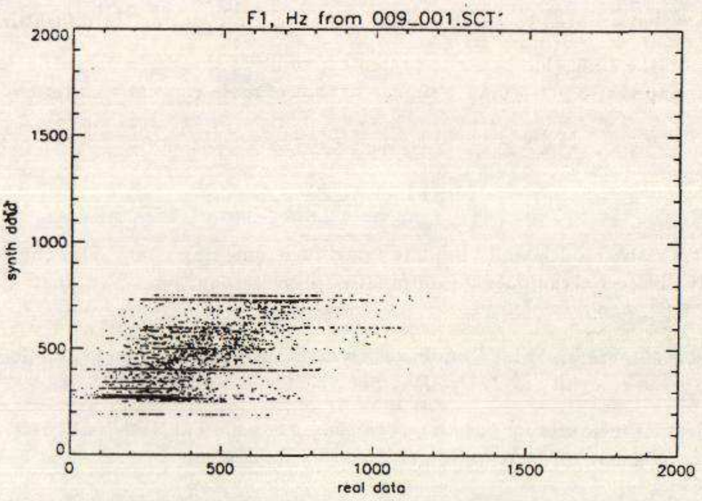

Figure 4: Scatter plot of first formant frequency for the first iteration of expt009.

done while it runs so slowly.

It may be that the amplitude-dependent deletion penalties reduce the distances generally. If so, a lower fixed pruning threshold might be appropriate and would run faster, but it is clearly impractical to have to optimise the pruning threshold for each experiment. Before it is worth trying to complete the adjustment of the matching parameters, it is necessary to devise a method for the adaptive control of pruning threshold which will equalise memory requirements under different experimental conditions. This should equalise the running times even though the range of scores changes, and also make the results more comparable. Care has to be taken, however, that it does not prevent the recogniser from generating many hypotheses during difficult parts of the input and few hypotheses during the easy bits.

5 CONCLUSIONS

The existing recogniser, despite its rather crude synthesis model using time-warped fixed-length segments, can be used successfully to investigate the basic problems of recognition by synthesis. Predictably, the problems uncovered so far concern the metric to be used when matching an indeterminate number of input spectral peaks to the three reference formants, and the control of the pruning threshold to obtain comparable results when parameters are changed from experiment to experiment.

Scatter plots, which enable the behaviour of the training process to be monitored for every frame in the training data, have proved to be a powerful means of investigating the difficulties of RbS.

It is hoped that, by the time this paper is presented, some of the basic difficulties of RbS will have been overcome so that a performance similar to that of more conventional recognisers can be reported.

REFERENCES

[1] J.S. Bridle. Stochastic models and template matching: some important relationships between two apparently different techniques for automatic speech recognition. *Proc. Inst. of Acoustics*, 1986.

[2] M.J. Russell, R.K. Moore, and M.J. Tomlinson. Dynamic programming and statistical modelling in automatic speech recognition. *J. Opl Res. Soc.*, 37(1):21–30, 1986.

[3] M.J. Hunt. Distance measures for speech recognition. Aeronautical Note NAE-AN-57, National Aeronautical Establishment, Ottawa, March 1989.

[4] M.J. Hunt. A robust formant-based speech spectrum comparison measure. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, ICASSP-85(paper 29.9):1117–1120, 1985.

[5] M.J. Hunt. Delayed decisions in speech recognition — the case of formants. Technical report, National Research Council of Canada, 1985.