Proceedings of the Institute of Acoustics

COMPARATIVE EXPERIMENTS IN RECOGNITION OF SUB-WORD UNITS BY HIDDEN MARKOV MODELS

Dave Miller and Peter Roach

University of Leeds, Leeds LS2 9JT

1. INTRODUCTION

Hidden Markov Models have become widely accepted as a powerful technique for recognition of units in speech ([1],[4]). However, while the general approach seems to be fairly generally agreed, there are various ways in which the technique may be implemented. We have been working on one component of the SYLK project ([3]), which aims to develop a syllable-based speech recognition system combining statistical and knowledge-based approaches to sub-word unit recognition, suitable as a front end for large-vocabulary, speaker-independent applications; in the course of this research we have explored a number of variations on the HMM theme. Effectively, one has choices in three areas: (1) in the type of speech unit that one tries to recognise and the data used for training and testing the models' performance; (2) one also has choices in the nature of the parameterisation of the acoustic signal, and (3) further choices in the nature of the model trained. This paper reports the effects of a number of such variables on HMM recognition performance.

In giving performance figures we will distinguish (as is now standard practice) between *% Correct* figures and *% Accurate;* the former counts the number of units of the "correct answer" that were successfully recognised and subtracts deletions and substitutions. This ignores any additional spurious items included by the system, while the latter measure gives a score that is reduced by such insertions.

2. CHOICE OF RECOGNITION UNIT

HMM's are not restricted to any particular type or size of unit and can be trained to recognise whole words or various sorts of sub-word units. There is a wide range of sub-word units available; the smallest is the *phonetic segment* or *phone.* A unit of a rather higher order is the *phoneme.* The fundamental difference between these two is that the phone is a physically observable unit that can be defined in acoustic terms; a phoneme is an abstract

linguistic unit that is not necessarily realized as a single discrete physical event. In the case of a syllable-initial prevocalic voiceless plosive (/p/, /t/ or /k/) the phoneme will usually be realised by a temporal sequence comprising a period of silence, then a brief but intense burst of noise, then noise excitation of the vocal tract (*aspiration*), then a vowel onset containing formant transitions. At the phone level each of these states would be likely to be treated as a separate unit, while phonemically they all collectively constitute the realization of the phoneme. Another example of the mismatch between phones and phonemes is that one phoneme may be realized simultaneously with another, whereas such overlapping is not usually admitted in the case of phones. For example, the nasal consonant phoneme /n/ in a word like 'daunting' may be realized in the form of nasalization of the preceding vowel: the word is phonemically /dɔ:ntɪŋ/ but phonetically (i.e. realized in phones as) [dɔ̃:tɪŋ]. Phone-level analysis may recognise a distinct (nasalized vowel) phone in such a case. Consequently, the phoneme is subject to a great deal of contextual variability which makes it a difficult unit to work with in recognition.

The set of phonemes is a fixed set determined by the theoretical principles that guide the analysis, just as the set of letters of the alphabet is fixed. But phones may be more fully or less fully specified, and phone-level recognition may be required to work with a large set of finely-discriminated units, or a small set of broad classes. We have previously worked with broad class phone recognition using "phonetic alphabets" of around 6 to 8 symbols ([7],[5]).

There are many alternatives to phone- or phoneme-level unit recognition. HMM's work well on whole-word recognition, though the limitations on vocabulary size are a problem as for all other approaches to whole-word recognition. The syllable is a unit of great importance in phonology, and the SYLK project uses the syllable as its principal recognition unit. However, for purely statistical modelling the number of syllables occurring in English (over 10,000) makes this a difficult unit to work with. Other contenders as possible units are demisyllables, diphones and triphones.

We have chosen to work with a non-standard recognition unit intermediate between the phone and the syllable, which we call the SYLKunit. We adopted the SYLKunit principally because the SYLK project aims to recognise speech in terms of a syllabic coding and we wished to explore the possibility of generating initial hypotheses about syllabic constituents for subsequent refinement by a knowledge-based system ([3]). It is conventional to divide the syllable up into *Onset* (any consonant(s) occurring at the beginning of the syllable), *Peak* (the

vowel at the centre of the syllable) and *Coda* (any consonant(s) at the end of the syllable). These three objects are known as *Syllable Constituents* ; there are approximately 60 Onsets in English, 20 Peaks and 120 Codas, though of course only a small number of the possible combinations of Onset, Peak and Coda are found in English. We decided that this was too large a set of units to recognise and, following Allerhand ([2]), produced a set of broadly classified syllable constituents (SYLKunits) consisting of 30 Onsets, 1 Peak and 60 Codas; as an example, the SYLKunit that we symbolise as STG comprises the following Onsets (given in phonemic transcription): /spr/, /str/, /skr/, /spl/, /skl/. Since no data currently exists that has been transcribed in appropriate symbols (what we call SYLKsymbols), we have devised an algorithm (based on the *Maximal Onsets Principle*) for re-coding TIMIT data in this form ([6]).

In addition to developing SYLKunit recognition we have worked with units of two other levels, mainly for comparative purposes since we wish to make comparisons with results obtained on similar data elsewhere. In all cases we have based the training and testing on the American TIMIT speech database which has been phonetically transcribed by experts with the help of a semi-automatic transcription system. One is a the phone unit used for the TIMIT transcription; the second is a much less fully specified broadly-classified phone. The basic transcription system used on TIMIT is a very detailed acoustic-phonetic phone labelling: we have felt it necessary to dispense with some unnecessary detail such as the difference between the silence phases of /p/,/t/and /k/. By not counting confusions within such highly similar groups we reach a level of transcription known as Reduced TIMIT ([4]). The second level is known as Broad TIMIT, in which major categories of sound are collapsed together so that all vowels are identified simply as "Vowel", all fricatives as "Fricative" and so on.

The results described below are therefore based on phone-sized segments; these results are what we use as baseline measures in our current work on evaluating our level of success in recognising SYLKunits.

3. EXPERIMENTAL RESULTS ON PHONE RECOGNITION

In our early work on recognition in Leeds (referred to above) we used a very coarse parameterisation of the speech signal, based on a four-channel filter-bank. Since the start of the current project we have used a 32-channel filterbank implemented in software. We present

here the results achieved on the Reduced TIMIT labels. The HMM used in all cases was a simple 3-state straight-through continuous density type with diagonal covariance matrix.

It would have been impractical to use the entire contents of the TIMIT CD (comprising 4200 sentences spoken by 420 speakers), and we therefore used a subset made of 1030 sentences taken from Dialect Regions 1 and 7, discarding the "duplicate" sentences and any containing obvious transcription errors. Two sentences from each speaker were kept as test data, the remaining ones being used for training.

To provide a baseline measure we used the output of a 32-channel bark scale filter-bank (linear scaling) as our input vector, producing the following results:

	% Correct	% Accurate
Reduced TIMIT	45.1	32.6

We then looked at Broad Class results, by ignoring identification errors if they fell within the same phonetic class of sound:

	% Correct	% Accurate
Broad Class	61.5	51.3

Using log scaling we achieved:

	% Correct	% Accurate
Reduced TIMIT	49.6	36.6
Broad Class	70.2	55.8

COMPARATIVE EXPERIMENTS WITH HMM'S

Work at RSRE on the ARM recogniser has shown good results using a cosine transform ([8]). We used a similar algorithm to transform our input vector:

12-Coefficient Cosine Transform

	% Correct	% Accurate
Reduced TIMIT	52.9	38.0
Broad Class	69.9	57.4

We also tried using a smaller number of coefficients:

8-Coefficient Cosine Transform

	% Correct	% Accurate
Reduced TIMIT	51.6	36.0

We next included the difference between the vector two forward and the vector two past and this, combined with the input feature vector, gave a vector of length 24:

12-Coefficient + Difference Vector

	% Correct	% Accurate
Reduced TIMIT	53.1	45.8
Broad Class	71.6	63.6

We then added a bigram grammar computed from the training set:

	% Correct	% Accurate
Reduced TIMIT	56.7	51.7

The above experiments were on male data only; it has never been our intention to limit our recognition work to male data, but this was used in the first instance for the sake of comparability with the majority of other published work. However, we have now trained models on female and male data. On Reduced TIMIT symbols, and using models trained on female and male speech combined, we find:

	% Correct	% Accurate
Male data	55.7	50.7
Female data	55.5	50.9
Female+male data	56.1	50.2

6. CONCLUSIONS

By systematically varying a number of conditions while keeping constant the training and testing data and the evaluation technique, we have arrived at a level of performance that we feel bears comparison with most results from other researchers working with similar data under similar conditions. The only performance figures we have found that significantly improve on ours are those quoted by Kai-Fu Lee et al ([4]): they claim 64% Correct and 53.2% Accurate; one difference is that the glottal stop, which occurs surprisingly frequently in English, is ignored in their recognition but treated as a phone to be recognised in ours. We have now to solve the problem of evaluating which of the units we have worked with will be most effective in generating initial hypotheses for the knowledge-based component of SYLK, and which will give the closest approach to successful lexical retrieval.

7. REFERENCES

[1] W.A.AINSWORTH, *Speech Recognition by Machine*, IEE. (1988)

[2] M.ALLERHAND, *Knowledge-Based Speech Pattern Recognition*, Kogan Page. (1987)

[3] P.D.GREEN, A.J.SIMONS and P.J.ROACH, 'The SYLK Project: foundations and overview', *Proceedings of the I.O.A.*, vol.12.10, pp.249-258. (1990)

[4] K-F.LEE, W-W.HON and R.REDDY, 'An overview of the SPHINX speech recognition system', *IEEE Trans. A.S.S.P., vol.38.1, pp. 35-45.*(1990)

[5] D.MILLER and S.ISARD, 'Aligning speech with text', *Proceedings of the I.O.A.*, vol.6.4, pp. 255-260. (1984)

[6] P.J.ROACH, D.MILLER, P.D.GREEN and A.J.SIMONS, 'The SYLK Project: syllable structures as a basis for evidential reasoning with phonetic knowledge', to appear in *Proceedings of the XIII International Congress of Phonetic Sciences.* (1991)

[7] P.J.ROACH, H.N.ROACH, A.M.DEW and P.ROWLANDS, 'Phonetic analysis and the automatic segmentation and labelling of speech sounds', *Journal of the International Phonetic Association*, vol. 20.1, pp. 15-21. (1990)

[8] M.J.RUSSELL, K.PONTING and S.M.PEELING, 'The Armada speech recognition system', *Proc. Voice Systems Worldwide.*(1990)