THE SYLK PROJECT: SYLLABLE STRUCTURES AS A BASIS FOR EVIDENTIAL REASONING WITH PHONETIC KNOWLEDGE

P.J.Roach*, D.Miller*, P.D.Green+ and A.J.Simons+

*University of Leeds
Leeds LS2 9JT, U.K.

+University of Sheffield
Sheffield S10 2TN, U.K.

ABSTRACT

This paper reports on work being done on the SYLK project, funded by the UK IEATP programme (project no. 1067): this is aimed at developing 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. Hidden Markov Models are used to construct initial hypotheses for the knowledge-based component; encouraging results in recognising different sub-word units are presented.

1. INTRODUCTION

The sub-word unit on which SYLK is based is the syllable: the acronym stands for '*S*tatistical *Syl*labic *K*nowledge'. An overview of the whole project is given in [5]. The arguments in favour of syllable-based recognition are well-known ([1]): the principal reason for choosing the syllable (and this is true to a lesser extent of the demisyllable and triphone units) is that much of the allophonic variation found in phonemes can be explained in terms of the syllabic position in which they occur. An example is the difference between voiced /r/ and its voiceless allophone [ɻ] found after /p t k/ in words such as 'pray', 'tray', 'cray': a phoneme-based recogniser trained to recognise /r/ would need to be trained to recognise voiced and voiceless allophones separately, whereas a system trained to recognise syllable onsets of the form `voiceless stop, r` would not need to be given variants: a voiceless /r/ is simply a normal property of syllables beginning in this way.

The motivation for the combined statistical and knowledge-based approach is that recognition by statistical model alone seems to work very well for the majority of straightforward instances of the units being recognised, though it is critically dependent on the initial training data; knowledge-based systems, on the other hand, have the ability to make use of multiple sources of knowledge to refine hypotheses at more and more detailed levels, but risk becoming fatally derailed if the initial hypotheses with which they start are incorrect. The ideal strategy therefore seems to us to be one which embodies a statistical component for making initial hypotheses, and a knowledge component for hypothesis refinement. In this approach, it is more important for the initial hypothesis not to be wrong than for it to be exactly right in full detail.

This paper is chiefly concerned with the initial, statistically-based part of the system, this being the one which has been most fully developed at the present time. In the full SYLK system, the lattice of SYLKsymbols provided from the first pass is used to instantiate (independently) hypotheses about the structure of each syllable in the utterance, centred on its peak. Allowed syllable structures, and their interrelationships, are made explicit by an object-oriented *Syllable Model;* further processing is based around the application of 'refinement tests' to the syllable structure hypotheses ([2]).

2. CHOICE OF UNIT FOR INITIAL HYPOTHESIS CONSTRUCTION

For large-vocabulary speech recognition, the most convenient form of output from the front-end is a *phoneme lattice* allowing subsequent lexical access from dictionary entries coded in terms of phonemes (though other lexical access techniques can be used). For the reasons explained above, however, we prefer not to work with phonemes as our recognition unit within the front-end: instead we envisage that the final stage in our front-end processing will be to recover a phonemic transcription from the syllable-based, allophonic explanation which SYLK will produce. Although our explanation unit is the syllable, there is no reason why we should not build initial hypotheses on the basis of phoneme-sized units if they can be reliably recognised. We may, for example, segment and label the speech signal in terms of acoustic phonetic units, where all major allophones of the phonemes are identified in a context-free manner. Alternatively, we may choose to identify phonetic segments that are members of a much smaller set: such broad phonetic categories (often based on manner of articulation, comprising categories like plosive, fricative, vowel, nasal) are likely to give more robust recognition (see [8],[10]). Another possibility is to attempt to recognise units above the level of the phonetic or phonemic segment. It is generally agreed that the number of syllables used in English exceeds 10,000, and to develop statistical models of all of these would not be computationally practical; consequently a unit smaller than the syllable may be best. Triphone modelling is used, for example, ARMADA ([11]); another unit which has its supporters is the demisyllable ([4],[12]).

For our purposes, bearing in mind that we are working towards decoding speech into fully-specified syllables at a later stage in the process, we prefer to make use of smaller units than demisyllables, but units which are explicitly tied to syllabic structure (which diphones and triphones are not). It is usual to view the syllable as composed of an optional ONSET, an obligatory PEAK (normally the vowel) and an optional CODA, each of which can be treated as independently recognisable objects ([1]). We believe there to be approximately 60 possible Onsets in English and about 120 Codas, while the number of Peaks is in the region of 20. Strangely, there appears to be no phonological term for referring in a generic way to Onsets, Peaks and Codas, and we are reduced to calling them Syllable Constituents. Although these units are potentially useful, we have chosen to work with units of the same size as Syllable Constituents but less fully specified. For example, we believe it to be unrealistic to expect a straightforward statistical recogniser to achieve speaker-independent, context-free discrimination of /spr/, /str/, /skr/, /spl/, /skl/, but we do think it feasible to aim to recognise the class of /spr/, /str/, /skr/, etc. If we bring together on acoustic grounds all highly-confusable Onsets and, separately, Codas into broader units, we reduce the set of Onsets to 30 and of Codas to 60. Again, no name exists for such units, but we have come to refer to them as SYLKunits ([9]).

3. EXPERIMENTS IN STATISTICAL

RECOGNITION OF SUB-WORD UNITS

We have been careful throughout this work to make use of widely-available and widely-used speech data and performance testing techniques so that our results should be comparable with research done elsewhere. Our original intention was to make use of a British English database as envisaged in the SCRIBE project, but delays in the production of this has obliged us to use instead the TIMIT corpus of American English. Since the total amount of data recorded on the current TIMIT CD-ROM disk is very large (4200 sentences spoken by 420 speakers), we have made use of a subset for training and testing purposes, based on the 1030 sentences collected from Dialect Regions 1 and 7; we discarded "duplicate" (SA) sentences and ones with obvious transcription errors. Two sentences from each speaker were kept as test data, the remained being used as training data. Female and male voices are being studied separately at present, and full results for the female voices are not yet available.

We have conducted a series of experiments in recognising sub-word units. Two different units were chosen, one a phoneme-sized unit based on the segments labelled in the TIMIT corpus, and the other the SYLKunit as described above. For the former, we trained models on every phonetic category. However, in its most detailed form, the TIMIT transcription distinguishes between *the silent portion* of /p/, /t/ and /k/, which is clearly not practical; by ignoring errors within such categories we effectively aimed at recognition at a level known as "reduced TIMIT" ([7]), roughly comparable in detail with phonemic representation. We have also tried "broad class" recognition of the same-sized unit.

Since no corpus annotated with SYLKsymbols was available, we had to produce our own. While some material in British English has been specially recorded and transcribed to give a full coverage to all possible Onsets and all possible Codas, our current use of American English and our need for large quantities of training data made it necessary to carry out an automatic re-coding of the TIMIT data into SYLKsymbols. This was done, making it possible to train HMM's for recognition of two different types of unit on the same recorded material. Since non-Peak SYLKunits are characterised as Onset or Coda, the re-coding required decisions about syllable boundaries; as is usual, such decision were based on the *Maximal Onsets* principal according to which all intervocalic consonants are assigned to the Onset of the following syllable if this does not violate phonotactic regularity.

It is essential to have a reliable and meaningful technique for scoring the recognition success rate. For work using TIMIT it has been usual to use the scoring technique developed at NIST for work on TIMIT, and we originally used this. We have recently adopted as our standard HMM software resource the HTK package developed at Cambridge University Electrical Engineering Department, and this contains a scoring technique that is similar to the NIST test. All our results given below were calculated by HTK scoring; we observe the standard scoring distinction between *correct* and *accurate* (where in the latter case, insertions cause a reduction of the score).

4. RESULTS

4.1 Recognition Scores

At the time of writing, the best scores we have achieved on the TIMIT test data are shown in Table 1 (data from male speakers only):

	Correct	Accurate
TIMIT LABELS.	56.6%	51.6%
SYLK-SYMBOLS.	67.9% (a) 60.8% (b)	53.5% (a) 57.7% (b)

Table 1: Recognition scores; (a) and (b) are from different HMM topologies.

It is important to compare these with results from elsewhere: the closest comparison we have been able to find is the context-independent phone recognition on TIMIT data reported in [7]: using male and female data, they reported 64% Correct and 53.2% Accurate. Glottal stops were ignored in their study, whereas we treat this as one of the phones to be recognised.

4.2 Comparative Evaluation: Phonetic Segments vs. SYLKunits

There remains an unsolved problem in interpreting these results: the two units studied are in some ways radically different from each other, and are not easily comparable. While excellent methods exist to compare two different attempts at recognition of a particular set of units in an utterance (e.g. [3]), what we have here is scores for units of different sizes and containing different amounts of information. We need to know which of the two units brings us *in principle* closest to successful word recognition. One way of doing this that we are currently investigating is first to discover which representation gives least uncertainty in word identification, using an approach based on [6]. We are using an on-line pronouncing dictionary of approximately 70,000 words and automatically re-coding the entries in SYLKsymbols and in TIMIT phonemic symbols. Each word, in both new representations, will then be checked against all the others to see how many other dictionary entries have identical coding, and the representation showing the smallest number of confusions will be shown to be the most favourable for word recognition. It should be remembered, however, that much might be gained from supplying the knowledge-based component of SYLK with both representations as partially independent sources of evidence.

5. REFERENCES

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

[2] BOUCHER, L.A., (1990), "Syllable-based hypothesis-refinement in SYLK", *Proc. I.O.A. 10.1*

[3] COX, S.J. (1988), "The Gillick Test - a method for comparing two speech recognisers tested on the same data", *Memorandum 4136*, RSRE Malvern.

[4] FUJIMURA, O. (1976), "Speech as concatenated demisyllables and affixes", *J.Acoust.Soc.Am.*, vol.59, p.55.

[5] GREEN, P.D., SIMONS, A.J. and ROACH, P.J. (1990), "SYLK Project foundations and overview", *Proc. I.O.A.*, vol. 10.1.

[6] HUTTENLOCHER, D.P. and ZUE, V.W. (1984), "A model of lexical access based on partial phonetic information", *Proc.ICASSP-84.*

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

[8] MILLER, D. and ISARD, S. (1984), "Aligning speech with text", *Proceedings of the Institute of Acoustics*, vol.6.4, pp.255-260.

[9] ROACH, P.J. (1990), "Phonemic transcription conventions and speech corpus design", *SYLK Working Paper No.7.*

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

[11] RUSSELL, M.J. et al (1990), "The ARM continuous speech recognition system", *Proc.ICASSP-90*, Vol.1, Paper S2.8, pp. 69-72.

[12] WEIGEL,W. (1990), "Continuous speech recognition with vowel-context-independent H.M.M.'s for demisyllables", *Proc.ICSLP-90*, Kobe, Japan, vol.2, pp.701-704.