Proceedings of the Institute of Acoustics

EXPERIMENTS WITH THE SYLK SPEECH RECOGNITION SYSTEM

P.D. Green, L.A. Boucher and N.R. Kew

SPLASH, Department of Computer Science, University of Sheffield

1. INTRODUCTION

The aim of the SYLK project is to develop a speech recognition front-end which combines statistical and knowledge-based processing, to mutual benefit. The project, located at Sheffield University and Leeds University, was funded by the UK IEATP programme from May 1989 to March 1992. The SYLK speech recognition approach is unusual in two ways:

> **The syllable, rather than the phone or the word, is chosen as the 'explanation unit'**; i.e. SYLK interprets a spoken sentence as a sequence of syllables, an hypothesises possible structures for each syllable. The motivation for the choice of the syllable is that much allophonic variation is conditioned by syllable position. The notion of syllable-based speech recognition is, of course, not new: a variety of approaches have been tried, for instance Allerhand [1], De Mori [2], Weigel [3]. The SYLK acronym stands for 'Statistical SYllabic Knowledge'.
>
> **Recognition proceeds in two stages: a conventional 'first-pass', followed by the application of 'refinement test'** (see figure 1). The first pass is based on Hidden Markov Models for syllable structure components; see section 2. Refinement tests provide a mechanism for using evidence (such as formant transitions) which the first pass cannot make explicit. Each refinement test attempts to capture some piece of phonetic knowledge and use it to improve the first-pass results. Refinement tests are trained and deployed within an evidential reasoning scheme. The result of applying a refinement test is to change SYLK's confidence in alternative hypotheses.

In the course of SYLK development, it became clear that existing assessment techniques were inadequate for a system which operates at multiple recognition levels. A companion paper (Kew et al [4]) deals with the **information-theoretic assessment technique** we developed in response to this problem. Work on the SYLK front-end is reported in Roach et al [5]. See Boucher [6] for details of the evidential reasoning scheme. Green et al [7] gives a more detailed account.

2. THE FIRST PASS

The SYLK first-pass is a conventional continuous-density Hidden Markov Model recogniser developed using the HTK toolkit (Young [8]). Since the aim is to attempt to find the likely syllabic structure of the input sentence, HMMs are trained for a set of **'SYLK symbols'**: allowed syllable onset, peak and coda types in English. There are about 20 such onsets, 20 peaks and 60 codas: table 1 ives examples. Training such models requires an appropriately-annotated database. Since no such resource exists, we derive training annotations by applying an algorithm based on the 'maximal-onset principle' to the conventional phonetic annotation of the TIMIT database, accepting that this procedure will inevitably introduce some syllabification errors. The training set comprises some 600 utterances from male speakers in the TIMIT dr1 and dr7 dialect regions. The test set is around 250 dr1

and dr7 utterances.

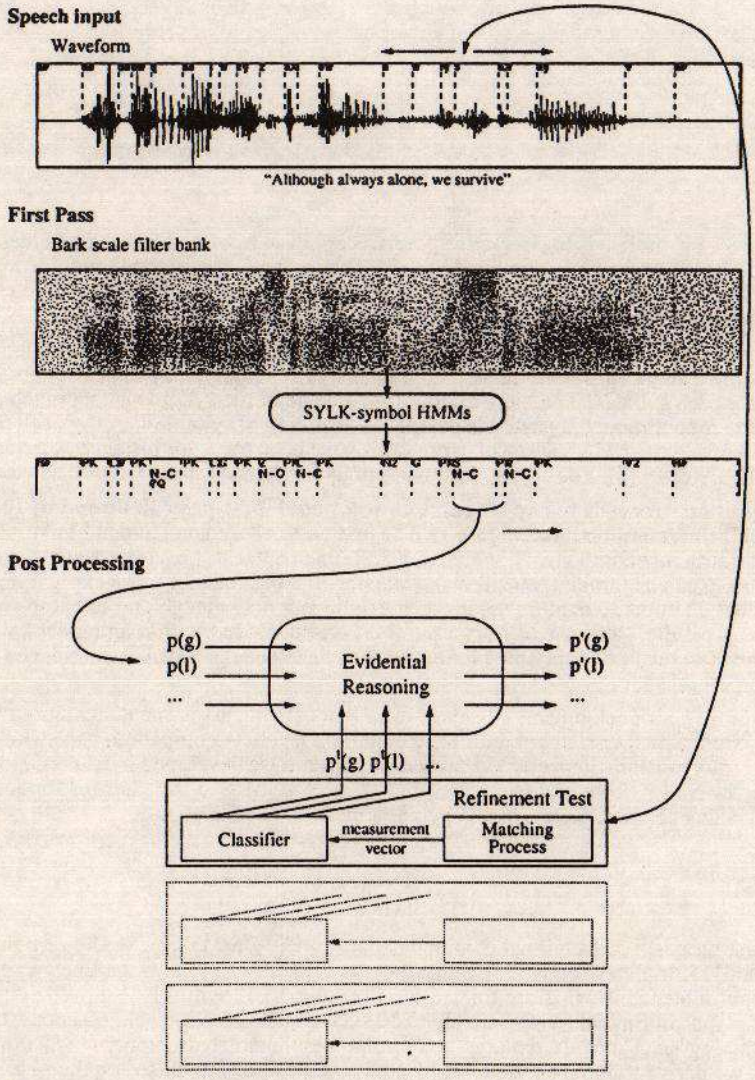


Figure. 1Recognition in SYLK

The representation used as input to the HMM recogniser is a 32-channel bark-scale spectrogram

Table 1: Examples of SYLK symbols

SYLK symbol	represents
T	voiceless-stop-onset /p- t- k-/
TL	voiceless-stop-liquid-onset /pr- pl- tr- kl- kr-/
STL	fricative-stop-liquid-onset /spr- spl- str- skr- skl-/
T2	voiceless-stop-coda /-p -t -k/
ST2	fricative-stop-coda /-sp -st -sk -st/

(the Audlab 'm32' filters), subjected to a cosine transform. The HMM topology is 3-state, 'straight through'. The overall performance of the first pass (Viterbi path) on the SYLK test set is, in summary

%Correct=66.2, Accuracy=54.5, RIT=0.594

Here **%Correct** and **Accuracy** are the convention NIST scoring measures and **RIT (Relative Information Transmitted)** is a measure of the proportion of the information required to construct the correct labelling that is present in the recogniser output (see Kew [4]).

In our experiments we have been concerned only with refining syllable onsets. Therefore we make comparisons against the first-pass performance on onsets only. Since the HMMs perform better for syllable peaks and because the process used in scoring is based around aligning peaks, the onset-only performance is considerably worse:

%Correct=41.3, Accuracy=38.2, RIT=0.279

The refinement process for a given syllable starts from an initial estimate of the probabilities of the various onsets. These probabilities may be obtained from a confusion matrix of first-pass errors or, more sensitively, by a technique which re-applies the first-pass HMMs, following the suggestion of McKelvie and McInnes [9]. Each interval between two peaks is considered as a separate utterance consisting of coda-onset and a lattice of possible HMM matches, with appropriate probabilities, is obtained. Taking the highest-scoring label in the lattice rather than on the Viterbi path, the first-pass performance on onsets make the performance slightly worse:

%Correct=40.28, RIT=0.259

3. APPLYING REFINEMENT TESTS

To add knowledge to SYLK, it is necessary to write a matching process M, which takes a syllable structure hypothesis H and returns a measurement vector v, in which alternative explanations for H should be statistically distinguishable. SYLK provides facilities for building a refinement test around M: in training, v is used as the basis for a statistical classifier which discriminates between the alternative labels for H, or its refinements. The system provides the choice between a Gaussian classifier or a neural-net (Multi-Layer Perceptron). When the test is applied, probability estimates from this classifier are combined with existing probabilities by Bayesian updating. We are also working on the use of neural nets for this evidence-combination to avoid the Bayesian assumption of independence of evidence.

EXPERIMENTS WITH THE SYLK SPEECH RECOGNITION SYSTEM

The SYLK architecture is intended to make the addition of knowledge easy and flexible:

(1) **In designing M** there is no restriction on:

what **representations** of the speech evidence are considered,
what **time segment(s)** are examined,
what **kind of processing** is performed,
what **programming language** is used.

(2) **In building the test**, the user can choose

what **level of refinement** to operate at,
how to **partition the possible refinements.**

(3) SYLK provides a consistent **mouse-and-menu driven interface** to the designing, training and assessment of refinement tests. Figure 2 illustrates some of this functionality in the training of the Sonorant Energy test (see below).

A refinement test may revise the probabilities ascribed to different labels in an existing recognition (a **'revision step'**), for instance

{ (0.4 T) (0.4 TA) (0.2 D) } -> { (0.6 T) (0.3 TA) (0.1 D) }

or it may provide estimates of alternative label probabilities at a finer level (a **'specialisation step)**, for instance

D -> { 0.8 /b/ 0.1 /d/ 0.1 /g/ }

For revision steps, we can quote NIST and RIT figures after the test is applied. For specialisation steps, we compare the RIT after applying the test with the maximum possible RIT, which would be obtained if the test performed perfectly:

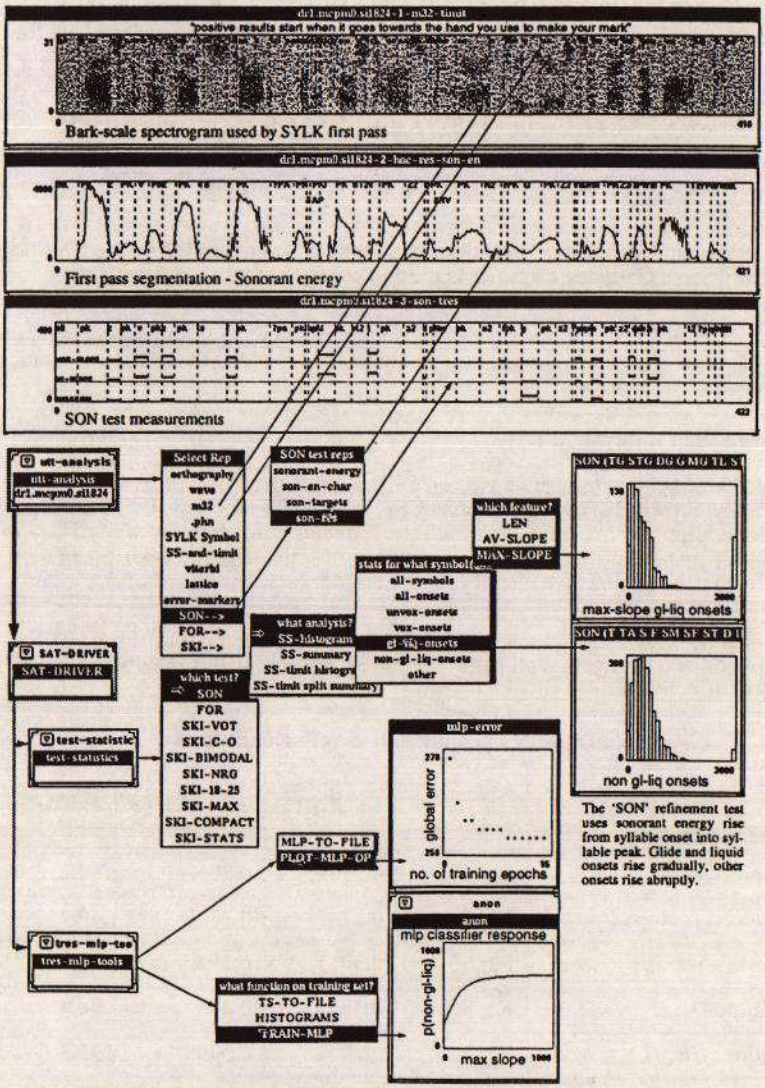
$$Itrans = \frac{(ActualRIT - PriorRIT)}{(IdealRIT - PriorRIT)} * 100$$

4. RESULTS and DISCUSSION

4.1 Sonorant Energy Test

Allerhand [1] suggested using the rise in sonorant energy from syllable onset into peak to distinguish abrupt onsets (rapid rise) from continuous onsets (gradual rise). In our implementation of this revision step, sonorant energy is derived from the 'm32' spectrogram, summing over the range 150Hz to 1.3KHz and taking logs. We use a technique developed from our earlier 'speech sketch' work (Green et al [10]) to describe the sonorant energy trace in terms of peaks and dips, find the vowel peak and measure its maximum slope.

EXPERIMENTS WITH THE SYLK SPEECH RECOGNITION SYSTEM



The 'SON' refinement test uses sonorant energy rise from syllable onset into syllable peak. Glide and liquid onsets rise gradually, other onsets rise abruptly.

Figure 2: SYLK analysis tool (SAT); allows:
Display of speech representations for selected utterances, Training of refinement test classifiers, Analysis of refinement test statistics.

Experiments with our statistical toolkit (histograms and scatter plots) showed that the most useful SON discrimination was between liquid-and-glide-onsets (SYLK symbols TG STG DG G MG TL STL FR DL and L) and the rest. This was the basis for classifiers developed for the SON test.

Using a **Gaussian Classifier**, we obtain the following:

%-correct 40.43 (from 40.28) Actual-RIT 0.2587 (from 0.2585)

Using an **MLP** classifier,

%-correct 39.79 Actual-RIT 0.2586

So this test makes little difference: a small but insignificant (as measured by the 'Gillick Test' [11]) improvement when a Gaussian classifier is used (4 changes of label in the right direction, 2 in the wrong direction out of a total of 835), a small deterioration with an MLP.

This is perhaps to be expected: the maximum slope measurement is something which the HMM-based first pass can effectively already use: it should correspond to the expected duration statistics for the appropriate model states.

4.2 Formant-based refinement tests

Formant-based refinement tests are based on a local implementation of Crowe's 'generalised centroid' algorithm (Crowe [12]), which estimates F1, F2 and F3 every 4ms. Using our 'speech sketch' technique we address the question 'Where do the formants come from as they go into the vowel?', with the aim of making distinctions based on place-of-articulation. Rather than taking measurements at a fixed point in time, we identify the start of a formant movement from a cartoonised representation.

Formant-based tests are used to make various specialisation steps below the SYLK symbol level. All the distinctions were made with MLP classifiers. We summarise results in table 2: recall that the **prior RIT is 0.2585.**

Table 2: Formant-based refinement tests

Distinction	Measures Used	Ideal-RIT	Actual-RIT	%-I-trans
voiced stops (/b,d,g/)	F1, F2, F3	0.3177	0.306	88.2
voiceless stops (/p,t,k/)	F1, F2, F3	0.3741	0.3391	69.7
nasals (/m,n/)	F2, F3	0.3051	0.2920	71.9
liquids (/l,r/)	F2, F3	0.3053	0.298	84.4
glides (/w,j/)	F2	0.3040	0.3005	92.3

The formant-based tests perform rather well, especially considering that our implementation of the Crowe algorithm is prone to mistakes in rapidly-changing segments such as plosive burst-follow-

ing vowel transitions.

Another obvious improvement is to include information about the following vowel, which is known to affect glide transitions for instance. We intend to do this by using the speech-sketch software to identify vowel formant targets, which we will include as additional test measures. Finally, we intend to try training separately over different onset clusters (i.e. different SYLK symbols) which end with the same distinction (e.g. /r/, /tr/, /str/).

4. 3 Plosive Discrimination Tests

We have experimented with various time and frequency-domain tests based on the work of Stella O'Brien [13, 14]. These specialisation steps are intended to determine place of articulation in singleton plosives;i.e.

D -> {b,d,g} (labial, alveolar, velar)

T -> {p,t,k} (labial, alveolar, velar)

The implementation is largely based on the **'Voiceless Speech Sketch'** (Kew [15,16]). Briefly, an FFT 'snapshot' of a burst is taken at the onset of the HF peak, smoothed, piecewise-linearised, and parsed for meaningful 'objects'.

We note that these tests are only valid for singleton plosives. This presents a problem in SYLK, due to the imperfect syllabification which may erroneously identify a singleton as part of a blend, or vice versa. An instance of this which has been extensively investigated is the plosive preceded by a nasal; for example the word "and" may be given as N-coda/D-onset, and coarticulation causes a /d/-profile to resemble a singleton /b/. This causes both recognition errors and (fundamentally) degraded training. The cumulative effect of such phenomena is that the tests are substantially weakened.

Characterising the Burst Energy Profile. One simple procedure is to bin the burst energy into 5 broad frequency bands (0-->1, 1-->2, 2-->3, 3-->5, 5-->8KHz) and train an MLP to make the 6-way distinction between /b,d,g,p,t and k/ given these measures. The test based on this MLP performs as follows:

Ideal-RIT 0.394 Actual-RIT 0.357 %-I-trans 72.6

In an attempt to characterise the energy profile in a more general manner, we measured the frequencies and maximum energy values of the two highest peaks in the profile. The aim here was to capture a variety of characteristic place-of-articulation behaviour, for example:

where the energy is concentrated in the LF range (bilabial),

where there is very high HF energy (aspiration; especially alveolar),

where energy is concentrated in particular peaks (duplicating "bimodal").

An MLP was trained to make the 6-way plosive distinction using the two peak frequency measures. Its performance was as follows:

Ideal-RIT 0.394 Actual-RIT 0.354 %-I-trans 72.1

'Compact and Diffuse' Energy Profiles. The burst energy profile in the range 800-3000Hz is said to be 'Diffuse Rising' for Alveolars, 'Compact' (dominated by a single peak) for Velars and' Diffuse Falling' for bilabials. To measure this, we fit a least-squares line to the data, and measures mean, gradient and variance. Thus {Alveolar -> high mean, rising}; {Bilabial -> low mean, flat or falling}. A compact peak (being a very poor fit to a straight line) will have a high variance. An

MLP classifier was trained to perform the labial-alveolar-velar distinction using the mean, gradient and variance measurements. Its performance was as follows:

Ideal-RIT 0.389 Actual-RIT 0.352 %-I-trans 71.6

In an attempt to make the compact property for velars explicit, we identified and measured the largest energy peak in the 800-3000Hz range. Any other peaks, however small, were also measured, and the ratio of the largest to the sum total was measured. An MLP classifier was trained to perform a velar v the rest distinction using this ratio. Its performance was as follows:

Ideal-RIT 0.419 Actual-RIT 0.338 %-I-trans 49.4

In summary, the plosive tests show some discrimination, but are generally weaker than the formant-based tests. It is clear that some of these tests are highly speaker-dependent, and need the addition of extra measurements to characterise the speaker.

5. REFERENCES

[1] MH ALLERHAND, 'Knowledge-based Speech Pattern Recognition', Kogan Page (1987).

[2] R De MORI, 'Computer Models of Speech using Fuzzy Algorithms', Advanced Applications in Pattern Recognition, series ed. M Nadler, Plenum, New York (1983).

[3] W WEIGEL, 'Continuous speech recognition with vowel-context-independent HMMs for demisyllables', ICASSP '90, Vol. 1, 69-72 (1990).

[4] NR KEW et al, 'An Information-theoretic approach to speech recognition assessment, with applications to SYLK', this volume.

[5] PJ ROACH, PD GREEN, DA MILLER and AJH SIMONS, 'The SYLK project: syllable structures as a basis for evidential reasoning with phonetic knowledge', Proc. XIIth Intl. Congr. of Phonetic Sciences, 4 (Aix-en-Provence, 1991), 482-485. (1991).

[6] LA BOUCHER, AJH. SIMONS and PD GREEN, 'Evidential reasoning and the combination of knowledge and statistical techniques in syllable-based speech recognition', in NATO ASI Series, vol. F75, 'Speech recognition and understanding: recent advances', P. Laface and R. De Mori, eds, Springer-Verlag Berlin Heidelberg 1992, p487-492 (1992).

[7] PD GREEN, LA BOUCHER, NR KEW and AJH SIMONS, 'The SYLK project - final report', University of Sheffield, Department of Computer Science, Research Report CS-92-18 (1992).

[8] YOUNG, SR, 'HTK Version 1.3: A Hidden Markov Model Toolkit', Cambridge University Engineering Dept. (1992).

[9] McKELVIE, D and McINNES, FR, 'Using entropy as a measure of Phoneme Lattice Quality', Proc. Eurospeech 1989, Paris (1989).

[10] PD GREEN, GJ BROWN, MP COOKE, MD CRAWFORD and AJH SIMONS, 'Bridging the Gap between Signals and Symbols in Speech Recognition', Advances in Speech, Hearing and Language Processing, ed. W.A. Ainsworth (JAI Press, 1990), 149-191.(1990).

[11] SJ COX, 'The Gillick Test: a method for comparing two speech recognisers tested on the same data', Memorandum 4136, Speech Research Unit, RSRE Malvern (1988).

[12] AS CROWE, 'Generalised Centroids: a new perspective on peak-picking and formant extraction', Proc. 7th Symposium of Fed. Acous. soc. Europe (SPEECH 88), WA Ainsworth and JN Holmes, eds, Inst. Acous. p 683-690 (1988).

[13] SM O'BRIEN, 'Evaluation of the Speech Knowledge Interface', HCC Report #29, LUTCHI, University of Loughborough, UK (1989).

[14] SM O'BRIEN, 'The Speech Knowledge Interface: Observations on the Identification of Plosives', HCC Report #41, LUTCHI, University of Loughborough, UK (1990).

Stella

[15] NR KEW and PD GREEN, 'A scheme for the use of syllabic knowledge in statistical speech recognition', Proc. 3rd Intl. Conf. on Speech Sci. and Tech., (Melbourne, 1990), 200-205 (1990).

[16] N.R. KEW 'Towards a voiceless speech sketch', Proc. Inst. of Acoust., 12(10), 373-380 (1990).