Proceedings of the Institute of Acoustics

THE SYLK PROJECT: FOUNDATIONS AND OVERVIEW

PD Green (1), AJH Simons (1) and PJ Roach (2)

(1) SPLASH, Department of Computer Science, University of Sheffield, UK.
(2) Department of Linguistics and Phonetics, University of Leeds, UK.

ABSTRACT

SYLK aims to combine statistical matching and phonetic knowledge in a front-end for automatic speech recognition. Processing in SYLK is based around the idea of refining hypotheses about syllable structures: the acronym stands for 'Statistical sYLlabic Knowledge'. Knowledge about syllable constituents (Onset, Peak, Coda ...) and their relationships is made explicit within an Object-Oriented Class inheritance lattice. In recognition, initial syllable structure hypotheses, centred around syllable nuclei, are derived from Hidden Markov Models for broad phonetic classes. These hypotheses are then refined by the application of 'tests', which are associated with one or more nodes in the syllable-constituent lattice. The execution of a test changes SYLK's confidence in the competing refinements, under a unifying evidential-reasoning scheme which uses statistical measures of the performance of the test over a training set.

In this paper we discuss the phonetic and methodological basis for SYLK, and give an overview of the system as it stands after 1 year's work. Companion papers deal with particular aspects of the project.

1. COMBINING KNOWLEDGE AND STATISTICS IN SPEECH RECOGNITION

It is commonly said (eg [Zue89], [Jung90], [Alle86]) that a combination of phonetic knowledge and statistical techniques should improve automatic speech recognition performance. However, there is no consensus about how one should go about combining knowledge and statistics. On the one hand, 'phonetic expert systems' (eg [Carb86], [Ster86]) have attempted to deploy rule-based knowledge-representation formalisms, perhaps with uncertainty-handling facilities ([Zue86], [Hata90]). It is now clear that such schemes do not transfer successfully to the ASR domain, both because of their representational inadequacy, and because they do not preserve either or both of the statistical virtues of trainability and admissibility. On the other hand, the idea of 'incorporating speech knowledge' into a statistical model by training on parameters appropriate to the phonetic class of the speech sounds being considered seems a weak use of the

phonetician's expertise.

In this overview, and its companion papers, we propose a more radical methodology for the development of what Allerhand [Alle86] has called a 'hybrid classifier'. Our basis is the notion of conditioning on a declarative model of syllable structure, rather than on an ignorance-based model like a Markov chain. The syllable model acts as our **'explanation unit'**: we interpret an utterance in terms of a sequence of instantiations of the model. We suggest a technique which enables speech knowledge to be expressed in trainable modules which are associated with the nodes of the syllable model. We make use of HMMs trained on syllable constituents to provide initial instantiations of the syllable model.

An outline of the recognition scheme in SYLK, and its motivation, is given in the section 2 below. The phonetic background leading to the choice of the syllable as an explanation unit is discussed in section 3. Section 4 deals with the computational embodiment of syllable structures. Section 5 deals with the initial HMM pass and the syllabic annotation of training databases. Section 6 is concerned with evidential reasoning with SYLK tests and Section 7 summarises the project position after around 1 year's work.

2. OUTLINE OF SYLK

SYLK (Statistical sYLlabic Knowledge) is an attempt to provide a framework within which phonetic knowledge can be deployed incrementally to explain acoustic evidence in terms of syllable structures. It has developed from work in the Alvey program at the Universities of Sheffield [Gree90], Leeds [Roac89] and Loughborough [OBri89], and has been influenced significantly by the work of Allerhand [Alle86]. SYLK is funded by the IEATP program.

In SYLK, knowledge is represented in terms of:

1. The **'SYLLABLE MODEL'**, a structure which expresses the phonological constraints that govern speech sound sequences, based on the syllable.

2. **'REFINEMENT TESTS'**, which relate syllable constituents to evidential representations. The application of a test is meant to improve the discrimination between alternative syllable structures. The performance of each test is assessed over a training corpus. There is no restriction on what evidence a test looks at, or what kind of processing it performs.

3. [Projected] **'STRATEGIC KNOWLEDGE'**, about which refinement tests to apply in which order in which circumstances.

In outline, recognition in SYLK works as follows:

1. An initial, HMM-based pass produces a lattice of **'SYLK Symbols'**: these are drawn from an alphabet of (phonetically) broad-class syllable constituent labels, corresponding to different kinds of syllable-onsets and syllable-rhymes (see section 5).

2. The SYLK -Symbol lattice is used to produce initial **'Syllable Structure Hypotheses (SSH's)'**. Each syllable nucleus in the lattice generates a **'Hypothesis-Set'** of SSH's which are in competition to explain the structure of this syllable. The SSH's will not be independent: some will be refinements of others. For instance, we might have a hypothesis-set containing 3 SSH's along the lines of

{
[SSH01: this syllable has a peak between t1 and t2 and a Voiced-Onset]
[SSH11: refinement-of SSH01: this syllable has a Continuous-Onset]
[SSH12: refinement-of SSH01: this syllable has an Abrupt-Onset]
....
....}

3. Each hypothesis-set is independently processed by invoking refinement tests under the control of a scheduling algorithm. On each cycle of this algorithm, some SSH, say h, is selected from the hypothesis-set and one of its refinement tests is applied. As a result, SYLK's confidence in the alternative refinements of h changes. For instance, we might apply a test which looks at the history of overall energy between onset and peak. This test might be applied to **SSH01** to improve the discrimination between **SSH11** and **SSH12**. Where necessary, new SSH's will be created and added to the hypothesis-set, for instance we might apply a presence-of-nasality test to **SSH12**, producing

[SSH121: refinement-of SSH12: the syllable has a nasal onset]
[SSH122: refinement-of SSH12: the syllable has a non-nasal onset]

Note that the performance of a test may effect SSH's other than the one over which it was invoked: there is no restriction that each test is associated with only one node in the Syllable Model. For instance, the presence-of-nasality test would also change SYLK's confidence in the onsets /sm.../ and /sn.../, which are

refinements of **unvoiced-onset**.

4. [Projected] Neighbouring hypothesis-sets compete to explain the evidence at their edges.

SYLK is meant to provide a context in which knowledge can be applied within a statistical regime: it is intended to overcome many of the objections to knowledge-based ASR:

1. Knowledge-based inference works best when a 'correct' line of reasoning is being expanded, but is notoriously weak without reliable seed hypotheses on which to build. In SYLK, the initial statistical pass provides these seeds, so that **the knowledge expressed in refinement tests is used to deepen and strengthen explanations which have a reasonable basis.** It is not necessary for the first pass to produce accurate syllable boundaries, or for every syllable nucleus it proposes to be correct, provided that no nucleus is missed.

2. **SYLK does not take hard-and-fast decisions,** in the manner of simple rule-based systems. Rather, the application of a test changes the relative confidence in alternative hypotheses, as indicated by the test's behaviour in training, but does not rule anything out. At any time in the development of an SSH, the probability measures attached to each SSH will reflect a judgment on the basis of the tests so far carried out.

3. **SYLK does not segment-and-then-recognise:** segmentation points in the initial lattice are used only loosely in specifying the time-region over which a test should be applied: tests are free to take evidence from where they wish, so that for instance a test on voiced onsets might also pay attention to the following vowel peak. This is in sharp contrast to the approach of SUMMIT [Zue89].

3. THE SYLLABLE AS AN EXPLANATION UNIT

The concept of the syllable has been important in phonology for at least two thousand years, though the status of the unit has varied with changes in fashion. The last few years have seen a considerable convergence of views in favour of the syllable among specialists in many areas of speech science. In theoretical phonology, for example, the approach of Chomsky and Halle [Chom68], which explicitly denied any theoretical status to the syllable, has given way to treatments in which the syllable has a central role, such as CV Phonology [Clem83] and Metrical Phonology [Selk82]. It should be mentioned that syllable grammars similar to those found in most recent treatments have

been around for some time (eg [Fudg69]), but have not always been given the importance they deserved.

Studies in speech production and perception seem to show growing evidence favouring syllable-based organisation. Mehler, Segui and Frauenfelder [Mehl81] argue from reaction-time experiments that syllable-sized units are accessed faster than phoneme-sized ones; and also find that pre-literate societies are more conscious of syllables than phonemes. Browman and Goldstein's X-ray microbeam data [Brow88] seem to associate syllable onsets comprising different types and numbers of consonants in precise temporal relationship ('C-centres') with the syllable nucleus, whereas syllable-final consonants behave in a less constrained manner. Studies of speech errors have, of course, for long been showing evidence of syllabic organisation.

Various kinds of syllable-based units have been proposed for speech recognition. These include the syllable [DeMo83, Chur83, Alle86], dissyllable [Sing88], demi-syllable [Rose81, Rusk81] and demi-syllable plus affixes [Fuji76]. Two overriding criteria determine the choice of unit size and structure. The unit must be sufficiently large to capture local phonetic variability, determined by the phonology of the language. However, too large a unit results in there being too many models to match against the data during recognition. There are approximately 10 000 regularly occurring syllables in the English language. If these are split at the nucleus, this gives approximately 800 initial, and 1 200 final demi-syllables. If affixes are treated separately from initial and final consonant clusters, the total number of units drops to around 1 600.

Following Church and Allerhand [Chur83, Alle86] we suggest that a considerable amount of context-dependent information, usually expressed as context-sensitive rules in the standard phone- or phoneme-based approach, can actually be rewritten in the form of a context-free syllable discrimination network. We adopt a simple syllable model, whose major constituents are:

```
Syllable --> [Onset] Rhyme
Rhyme    --> Peak [Coda]
```

but, because the sharing of syllable sub-structure is permitted and modelled in the network, the overall complexity is no greater than for demi-syllables. Onset and Coda clusters are progressively refined down to the level of 'SYLK Symbols', approximately a mid-class acoustic-phonetic labelling that treats Onset and Coda occurrences as potentially distinct. The meaning of the SYLK Symbol level of description is clarified in section 5.

A point often not considered enough by speech engineers is that phonological models do not necessarily translate into recogniser architectures. One example of this is the consistent failure of Distinctive Feature based recognisers, which confuse the linguist's phonological categories with the engineer's physical measurements. However, we are convinced that phonology does not appear 'out of the blue', rather phonological models are constructed on the basis of good phonetic evidence. A syllable structure model can be shown to reflect a systematic set of acoustic phonetic relations between syllable constituents and, to this extent, we are claiming a statistically-conditioned, phonetic basis for the syllable. In this respect, the syllable model plays the same role as the linear sequence of states on which HMMs are conditioned; except that the structure of our model is motivated by *a priori* phonological knowledge. We see the syllable as the 'explanation unit' over which much acoustic phonetic variability is predicated, especially the complex timing relationships of English.

4. THE OBJECT-ORIENTED SYLLABLE MODEL

The implementation vehicle for symbolic processing in SYLK is CLOS, the Common Lisp Object System [Keen88], running on SUN workstations. CLOS makes it easy to define classes to represent speech-concepts and allows multiple-inheritance to be used to represent the various relationships between these classes. The environment for SYLK includes an object-oriented interface with speech databases such as TIMIT [Fish87], and display facilities for speech-evidence representations based on multi-methods [Gree90b].

The Syllable Model can be viewed as a network of nodes connected in several planes. Amongst other things, nodes may represent phonetically salient objects including:

* **the syllable and its constituents:** onset, rhyme, peak, coda;

* **refinements of onset and coda:** abrupt, continuous, voiced, voiceless, down as far as consonant/acoustic segment clusters;

* **refinements of peak:** short, medium, long, extra-long, down as far as vowels and diphthongs;

* **acoustic segments:** silence, sibilance, vocoid, burst, aspiration, nasal, glottalisation;

* **acoustic objects:** formants characterised as peaks, dips; and aperiodic sources characterised by their envelopes, centres of

mass and first excited poles.

These nodes are linked in several planes. Each plane may be seen as a uniform perspective on some set of objects that may be reached by following links of the same name. The following links are represented:

* **constituents:** the time-ordered immediate constituents of syllables, or clusters; eg a ClosedRhyme has two constituents, Peak and Coda;

* **refinements:** the nearest nodes in hypothesis space, reached by applying one refinement test; eg an Onset has two possible refinements, VoicedOnset or VoicelessOnset;

* **components:** the concurrent frequency-ordered acoustic components of acoustic segments; eg a vowel-like peak has three significant (moving) formant-objects as its components;

* **subclasses:** the specialisations which inherit most of their description from the current node; eg [t^h] and [t'] are the specialisations of /t/ to be found in the onset and coda, respectively - this relation recaptures the notion of the phoneme, treating it as the superclass of its allophonic children.

The syllable model is described more thoroughly in [Simo90].

In addition to this scheme, a certain number of computational abstractions are modelled, including the notions of an **Utterance** (the object controlling access to data related to one single utterance), a **Collection**, which controls access to a number of utterances with specified properties, for use in training, a **Hypothesis**, which controls access to a particular SSH for some syllable in a given utterance, and a **Hypothesis Set**, which controls the processing of SSH's in competition to explain a given syllable.

5. PROVIDING SEED HYPOTHESES

To provide seed hypotheses for subsequent refinement, we use a statistical (Hidden Markov Model) approach, developed in the Leeds University Phonetics Laboratory as part of Alvey project MMI/053 [Roac89], for recognition of sub-word units. Originally the units recognised were phonetic segments identified with a small set of Broad Class labels (e.g. Plosive, Nasal, Vowel); it was shown that despite the constantly varying nature of the speech signal it is in general possible to segment and label

speech with an accuracy significantly better than chance [Roac89]. The original training and testing data comprised a corpus of English sentences and nonsense words recorded by equal numbers of female and male speakers and subsequently hand-labelled by expert phoneticians using the Broad Class labels. On test material, identification of the Vowel category was around 80% accurate, and for the purposes of building a hypothesis about syllabic structure in an unknown speech signal the identification of syllable peaks is clearly of primary importance. While in current work we are hoping vowel recognition will improve beyond the current level, we are also trying to refine the identification of other parts of the syllable.

In SYLK, we have chosen to work with a very much larger data set for training and testing: in the long term we plan to use the British SCRIBE database material [Hier90] when it is available in substantial quantity, but for the present we are using sections of the American TIMIT database [Fish87] in addition to the material we have recorded and labelled ourselves. We are currently experimenting with two different recognition tasks: in one, a much larger set of phonetic categories is used, based on the TIMIT symbol inventory. It is clearly unrealistic to expect the system to recognise some of the very subtle differences annotated, such as the difference between the silent hold portions of /p t k/, so a reduced set suggested by Robinson and Fallside [Robi90] is used, resulting in a symbol inventory of 41. The other approach is to attempt to recognise sub-word units larger than the phonetic segments referred to above: since the project as a whole is based on the idea of the syllable as an explanation unit it seems to us desirable to acknowledge this is our choice of unit for statistical modelling..

If we represent each onset and coda phonemically, and treat all vowels as belonging to a single category P (for Peak), this results in a total set of just under 200 units, comprising 65 onsets and 130 codas. However, many of these items are acoustically closely similar (e.g. '-enths' and '-engths'), while the difference between them carries virtually no functional load, so we feel justified in grouping them into what we might call "broadly classed syllable components". We have devised a notation system for these components (the **'SYLK Symbols'** of section 3 above). The set of SYLK Symbols comprises 22 onset types and 65 coda types [Roac90]. We have written software to convert strings of segmental phonetic symbols into SYLK Symbols and have begun experimenting with Hidden Markov Models trained on these syllabic units. We are thus able to compare recognition accuracy on the same test data using first segmental phonetic labels and then SYLK Symbols.

The American data that we are using at present diverges from British English in a number of ways, but for our purposes it is possible to overcome most of the problems: for example, syllables containing non-prevocalic /r/ are dealt with by incorporating the /r/ into the Peak. Vowel quality differences are irrelevant at the moment, since we are not trying to differentiate vowels at this early stage of the recognition process.

We must emphasise that the accuracy of recognition of phonetic detail in the first pass is not central to SYLK as a whole: the later refinement tests are where the fine phonetic and phonemic distinctions are made.

6. REFINING HYPOTHESES

To construct a refinement test, it is necessary to first define a **Matching Process**, whose input is a Syllable-Structure Hypothesis **H**. The matching process examines the evidence for h and produces a feature vector, whose elements may be numeric or symbolic. In the feature vector, the alternative refinements of **H**, $\{h_1\ h_2\ ..\}$, should be (statistically) distinguishable. In training, the Matching Process is used to produce **Test Statistics**, to allow the estimation of partial density functions for the h_i . When the test is invoked in recognition, the feature vector produced by the Matching Process for the SSH under consideration, together with the Test Statistics, are used to update the probabilities of the h_i. The updating scheme is a form of Dempster's rule [Shaf76]. Details of the evidential reasoning scheme are given in a companion paper [Bouc90]. The same matching process may be associated with more than one node in the syllable-refinement plane, in which case it will be conditioned separately for each node, to form tests which are separate, but always invoked together.

Thus a test implementor need only design the matching process, and there is no restriction on what evidence this process looks at, or what kind of processing it does. In practice, we expect many tests to follow the 'describe-and-compare' philosophy which we have previously advocated [Gree90]. However, we also anticipate, for instance, tests which invoke individual HMMs within selected regions of the utterance. We also expect that in addition to producing its feature vector, a test will leave other structures behind to aid in subsequent refinement processing, for instance evidence-descriptions (Speech Sketch elements) which it has computed, and timings of objects that have been matched.

7. PROGRESS SO FAR

Most of the first year of SYLK has been spent in setting up the computational environment described above. At the time of writing the situation is as follows.

The syllable model is complete and implemented,

We have reduced-TIMIT HMMs for the first pass, and are experimenting with HMMs for SYLK Symbols.

The refinement framework and the evidential reasoning scheme are in place.

We are in the process of training the first generation of refinement tests. These include characterisation of burst profiles and voice onset times [Kew90], and voiced-onset discrimination based on a description of formant movement between onset and peak.

REFERENCES

[Alle86] M H Allerhand (1986), 'A knowledge-based approach to speech pattern recognition', PhD Dissertation, Darwin College, Cambridge UK.
[Bouc90] LA Boucher and PD Green, 'Syllable-based Hypothesis-Refinement in SYLK', this volume.[Brow88] C P Browman and L Goldstein (1988), 'Some notes on syllable structure in articulatory phonology', Status Report on Speech Research 93/4, Haskins Laboratories.
[Carb86] N Carbonell et al, 1986, 'APHODEX: design and implementation of an acoustic-phonetic decoding expert system', Proc ICASSP-86, pp 1201-1204.
[Chom68] N Chomsky and M Halle (1968), The Sound Pattern of English, Harper and Row.
[Chur83] K W Church (1983), 'Phrase structure parsing: a method for taking advantage of allophonic constraints', PhD Dissertation, MIT.
[Clem83] G N Clements and S J Keyser (1983), 'CV Phonology' MIT.
[DeMo83] R de Mori (1983), Computer Models of Speech using Fuzzy Algorithms, Plenum.
[Fish87] W Fisher et al, 1987, 'An acoustic-phonetic database', JASA Suppl (A), 81, S92.
[Fudg69] E Fudge (1969), 'Syllables', J. Linguistics, Vol. 5.
[Fuji76] O Fujimura (1976), 'Syllables as concatenated demi-syllables and affixes', JASA, Vol. 59, p 55.