A SCHEME FOR THE USE OF SYLLABIC KNOWLEDGE IN STATISTICAL SPEECH RECOGNITION

N.R. Kew and P.D. Green

Department of Computer Science
University of Sheffield
Sheffield, England

ABSTRACT - We describe a new project, SYLK, which aims to combine statistical and knowledge-based approaches in a front-end for Automatic Speech Recognition. It is based on the syllable as an explanation unit. The processing comprises an HMM front-end, which is followed by an inferential reasoning system in which a series of individual refinement tests are applied to enhance the overall performance.

We then consider the task of plosive discrimination in the context of SYLK, and illustrate the flexibility of the system in its ability to encompass a variety of approaches. Some preliminary results demonstrate the utility of the syllabic approach.

INTRODUCTION

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 (Green et al, 1990), Leeds (Roach, 1989) and Loughborough (O'Brien, 1989), and has been influenced significantly by the work of Allerhand (1986). An overview of the project is given in Green, Simons and Roach (1990).

In this paper, we first give a brief outline of SYLK, presenting the syllable model, the probabilistic processing scheme, and the "refinement test" through which arbitrary local processing is incorporated. We then consider the problem of devising such tests, applied to the task of plosive discrimination. We describe a number of tests, and give some preliminary results which demonstrate the potential power of the syllabic approach.

AN OUTLINE OF THE SYLK PROJECT

Introduction

Recognition in SYLK is based on a structured syllable model, which expresses the phonological constraints that govern speech sound sequences. This acts as our 'explanation unit': we interpret an utterance in terms of a sequence of instantiations of the model. A probabilistic state of belief is maintained, in which different hypotheses compete to explain the acoustic evidence around a syllable nucleus in terms of the model. We make use of HMMs trained on syllable constituents to provide initial instantiations of the syllable model. Further processing comprises the application of "refinement tests", which provide a trainable framework for the incorporation of knowledge.

The Syllable Model

Following Church (1983) and Allerhand (1986) 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]

Onset and Coda clusters are progressively refined down to the level of "SYLK Symbols" [Roach], approximately a mid-class acoustic-phonetic labelling that treats Onset and Coda occurrences as potentially distinct. For example, SYLK symbol D denotes voiced plosive phonemes (/b/, /d/, /g/) in Onset position, whilst D2 denotes the corresponding coda. However, neither of these is used for a plosive in a consonant cluster; for example, a voiced plosive leading into a liquid is denoted DL.
The Syllable Model can be viewed as a network of nodes connected 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 principal links represented are:

* constituents - the time-ordered immediate constituents of syllables (as detailed above), or clusters.

* refinements. This plane may be seen as representing a structured set of phonetic classes appropriate to a constituent. Figure 1 shows the refinement plane for Onsets; we see that, for example, Onset has two possible refinements, VoicedOnset or VoicelessOnset.

The SYLK system uses an object-oriented implementation of the syllable model (Simons, 1990).

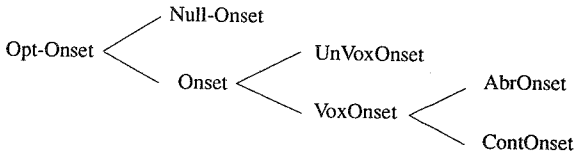

Figure 1. Refinement Plane for Onset in the Syllable Model (Principal Nodes).

A Probabilistic Framework for Recognition

The initial HMM pass over a speech signal is used primarily to detect peaks, which are then instantiated as syllable nuclei, and the HMM output is combined with a sonorant energy detector in this task. A second function of the HMMs is to produce initial onset and coda hypotheses, of which there are generally several candidates to explain the evidence. There may also be uncertainty regarding syllable boundary, in which case competing hypotheses will ascribe evidence to a coda or the following onset (e.g "cat sat"/"cat's at"). The competing hypotheses are ascribed probabilities by the HMM's. These are then incrementally updated by successive refinement tests, governed by a Bayesian inference machine (Boucher & Green, 1990), using a constrained form of Dempster's rule.

Refinement Tests and Low Level Processing

The mechanism by which phonetic knowledge and arbitrary low-level processing is incorporated in this framework is the Refinement Test (Kew, 1990). Briefly, this comprises a Process and a Training, and acts at a syllable model Node. The Process describes some aspect of a fragment of speech signal, returning a feature vector. The Training comprises a set of probability density functions, representing the various hypotheses between which the test discriminates, and by reference to which a feature vector is interpreted. A test is trained over one or more syllable model nodes, and a process may be used in several tests by training at different nodes, thus taking advantage of the syllabic structure. There is no restriction on the processing a test may perform, nor on the speech signal representations it may use.

The particular utility of the refinement test is that it provides a well-defined interface for low-level processing which is (a priori) totally arbitrary, in a statistically optimal Bayesian framework. The test developer can concentrate on the tasks of pattern recognition and feature extraction, without reference to the interpretation of his results.

A key feature of our approach is that a representation may be re-used in several tests, each of which is small and sharply focussed on a particular feature (and hence also highly trainable). This approach is demonstrated in the tests on a plosive profile described below, and contrasts with Ruske's otherwise similar syllable-based approach to plosive classification (Ruske, 1983). We note that our approach requires some care, as re-use of the same data may lead to erroneous results if different tests are not independent. We propose to avoid this by incorporating a measure of the orthogonality of different tests in our probabilistic updating rules.

PLOSIVE DISCRIMINATION TESTS

The plosive discrimination task is one which has been addressed in SYLK. We consider how tests may be devised in this task, and describe a number of those tests which have been implemented or are currently being considered.

Emulation of the Spectrogram Reader.

The Expert System approach adopted has been based on the study of plosive articulation by O'Brien (1990). Our use of this approach has concentrated on frequency-domain information in a plosive burst, which is examined for general trends and particular salient features. The principal mechanism for the use of such information in SYLK is a frequency-domain speech-sketch, based on parsing a smoothed short-time Fourier transform of the burst (Kew, 1990). Such tests include characterisation of:

* The profile in the region of 800-4000Hz as high/low, rising/falling, compact/diffuse. One test process returns basic statistics as a feature vector; another specifically looks for compactness.

* The onset of a region of high frequency noise energy is measured in a test described as "cut off".

* The distribution of energy in the profile is quantised in significant regions.This is seen as an alternative to cut-off, and may prove more robust.

* Large peaks in 1000-1800 and 3000-5000Hz and in a frequency ratio of approximately 3:1, strongly indicative of velar articulation.

* Small spectral prominences at about 1800 and 2500Hz in a rising profile, strongly indicative of alveolar articulation.

Use of Perceptual Experiments.

A number of parametric synthesis-based experiments have demonstrated the importance of particular aspects of the speech signal in plosive perception. Results such as those of Lisker & Abraham (1964) clearly suggest Voice Onset Time (VOT) as a suitable subject for a SYLK test. Similarly, experiments such as Datscheweit (1989) indicate the potential of a test based on Formant Transitions (although we note that these results appear not wholly to agree with the observations of O'Brien).

It should be noted, however, that these experiments avoid the variability of real speech, so we cannot expect these SYLK tests to show such clear distinctions as their laboratory counterparts.

Automatic Tests.

A third form of test process is the automatic characterisation of some speech signal representation. The use of a self-organising neural net to characterise burst profiles (and eventually other speech representations) is currently under investigation.

PRELIMINARY RESULTS

We are not yet able to present results within the full framework of SYLK. However, we discuss the interpretation of results, and quote results from some tests running in isolation. We use these to demonstrate the power of the syllabic approach. Our sample is a designated training set from dialect region 1 of the TIMIT database (Fisher et al, 1987).

Syllabification.

As the full SYLK environment is not yet available, our tests cannot be trained over the syllabic hypotheses. Instead we rewrite the TIMIT annotations as SYLK symbols. These are necessarily tentative, and a principle of maximal onsets is arbitrarily applied. Inevitably some codas will be represented as onsets in this scheme. We attempt to compensate for this by defining three datasets: Onset (definite), Coda and Intervocalic (V-P-V). The codas are unfortunately too few in number to give reliable results, so the efficacy of the syllable model can only be demonstrated by comparing behaviour of tests trained on onsets to equivalent results trained over all plosives.

Assessment of Test Performance.

The complete SYLK system may be assessed in a number of standard ways, such as percent accurate and correct. However, in the context of the probabilistic framework for representation of belief, it is not sensible to assess individual refinement tests by measures such as "percent correct". To see this, we consider a hypothetical example in which a refinement test makes a three-way distinction using a one-dimensional feature vector. Figure 2 shows feature vector distributions for the three classes; we see that a maximum likelihood selection will hardly ever select class b, although it occurs as often as the others. It is also clear that the merging of any two classes in a two-way distinction will entail considerable loss of information, and that the only sensible representation for interpreting the majority of inconclusive results is probabilistic.

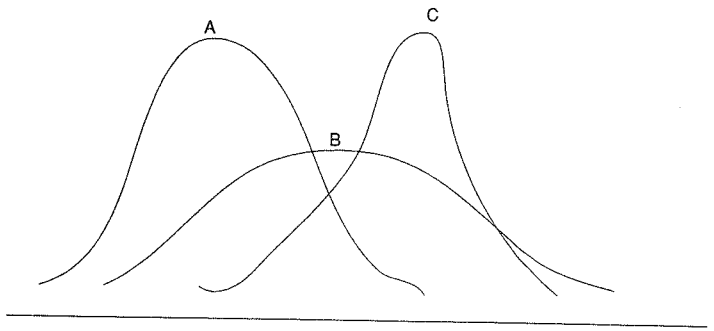

Figure 2. A Three-Way Distinction Not Adequately Represented by a Maximum Likelihood Selection Rule

Percent correct results in fact usuallyfavour a two-way test. An example is the test FD-STATS, which measures basic statistics concerning general trends in a burst profile, in a three dimensional feature vector. Training it over voiceless stops, we find that /k/ returns values approximately midway between /p/ and /t/, and best "correct" results are obtained by clustering /k/ and /t/. In this case, scores computed are 74% correct (Onsets). Making a 3-way distinction, this reduces to 66%, with only 33% of /k/'s being correctly identified. The corresponding non-syllabic figures are 73% and 53%, with the /k/'s scoring only 16%, which we consider inconclusive evidence for the syllabic training.

The behaviour of a refinement test may be represented either visually in a histogram or equivalent, as in the hypothetical case considered above, or by a measure of discrimination. For multi-dimensional feature vectors, only the latter is readily available. We choose a distance measure arising from our clustering algorithms to represent discrimination; in effect measuring the distance between classes normalised by the variability within a class.

Unfortunately, reliable discrimination results for a good selection of tests are unavailable at the time of writing; these will be presented at the Conference.

The discrimination of the test VOT over the dataset is:

0.166 (Non-Syllabic Training)
0.326 (Trained over Onsets)

representing a ratio of 2:1 in test performance. The remaining measures are 0.095 (Codas) and 0.262 (Intervocalic), suggesting that test VOT does not perform so well in coda position, although the dataset for codas is too small to draw reliable conclusions.

Visual Representation

Where a test process returns a single number (one dimensional feature vector), its behaviour may be represented visually in a histogram. The test VOT (Voice Onset Time) satisfies this constraint, and results are shown graphically in Figure 3. We observe that these results represent a training for VOT.

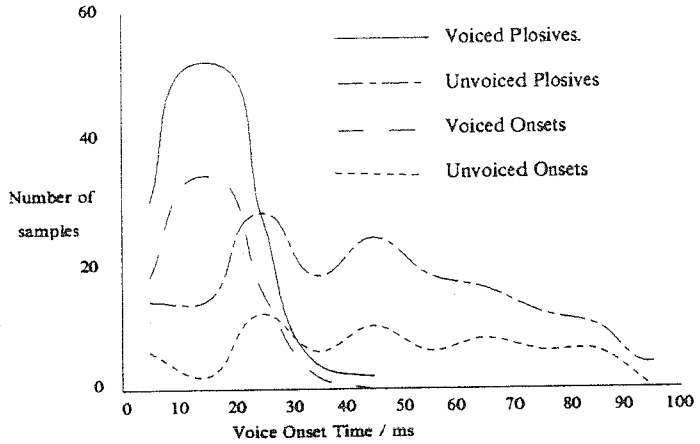

Figure 3. Distribution of Voice Onset Time for All Plosives / Onsets Alone.

It is clear from Figure 3 that the enhanced performance over syllable onsets may be attributed to the more extended distribution of voiceless VOT's in onset position.

REFERENCES

Allerhand, M.H. (1986) *A Knowledge-Based Approach to Speech Pattern Recognition*, PhD Dissertation, Darwin College, Cambridge UK.

Boucher, L.A., Green, P.D. (1990), *Syllable-Based Hypothesis-Refinement in SYLK*, Proc.IOA.

Church, K.W. (1983), *Phrase Structure Parsing: a Method for Taking Advantage of Allophonic Constraints*, PhD Dissertation, MIT.

Datscheweit, W. (1989) , *Quantitative Measurement of the Influence of Acoustic Cues on the Perception of Voiced Plosives*, Proc.Eurospeech.

Fisher, W. et al (1987), *An Acoustic-Phonetic Database*, JASA Suppl (A), 81, S92.

Green, P.D. et al (1990), *Bridging the Gap between Signals and Symbols in Speech Recognition*, in 'Advances in Speech, Hearing and Language Processing', WA Ainsworth (ed), JAI press, pp149-191.

Green, P.D., Simons, A.J.H., Roach, P.J. (1990) *The SYLK Project: Foundations and Overview*, Proc.IOA.

Kew, N.R. (1990) *Towards a Voiceless Speech Sketch*, Proc.IOA.

Lisker, L., Abraham, A. (1964) *A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements*, Word, 20, 384-422.

O'Brien, S.M. (1989) *Evaluation of the Speech Knowledge Interface*, HCC Report #29, LUTCHI, University of Loughborough.

O'Brien, S.M. (1990) *The Speech Knowledge Interface: Observations on the Identification of Plosives*, HCC Report #41, LUTCHI, University of Loughborough.

Roach, P.J. (1989) *Phonetic Feature Extraction by Automatic Segmentation and Labelling*, final report to SERC on Alvey project MMI053, Department of Linguistics and Phonetics, University of Leeds.

Roach, P.J. (1990) *Phonetic Transcription Conventions and Speech Corpus Design*, SYLK Working Paper #7, Dept. of Linguistics and Phonetics, University of Leeds.

Ruske, G. (1983), *On the Usage of Demisyllables in Automatic Speech Recognition*, Signal Processing II: Theories and Applications, H.W. Schlussler (ed), North Holland, pp419-422.

Simons, A.J.H. (1990) *Object-Oriented Syllable Structures*, SYLK Working Paper #2, Dept. of Computer Science, University of Sheffield.