A PROSODICALLY AND LEXICALLY CONSTRAINED APPROACH TO CONTINUOUS SPEECH RECOGNITION

J. R. Sholicar †, F. Fallside

Speech Laboratory, Engineering Department, University of Cambridge, U.K.

ABSTRACT

Psycholinguistic studies have indicated that prosodic cues play a vital role in human speech perception. The prosodic relationships which exist within an utterance are believed to provide fundamental cues for structuring the recognition process. However, in the majority of reported systems for the automatic recognition of continuous speech, prosodic cues are seldom used. In this paper, we review the evidence supporting the exploitation of prosodic cues, and discuss how such cues can be exploited within a machine recognition system to improve the segmental parsing strategy. A practical implementation is then proposed, in which prosodic structure is a major factor in the organisation of the recognition process. The architecture of the system is described and preliminary results relating to the current development of this system discussed.

INTRODUCTION

For many years the development of systems for the automatic recognition of continuous speech has primarily focussed upon the segmental (i.e. phonemic) description of the acoustic stream. Prosodic cues have been considered to be useful but optional extras which simply interfere with the process of segmental decoding: evident for example in the use of dynamic time warping, in which prosodic durational cues are removed. Within those systems in which prosodic cues are employed, prosodic constraints have merely provided additional restrictions by which to filter segmentally invoked lexical hypotheses (Vaissiere 1985).

There is however significant experimental evidence from studies in psycholinguistics and speech pathology of the importance of prosodic cues in providing structural organisation for human speech communication. Freeman (1982) for example claims that prosody is intrinsic and critical in both perception and production of speech, and that it functions as the foundation or structural support for the organisation of speech communication. Within this working paper, we aim to re-assess the bias placed on segmental decoding within machine recognition systems, investigating how the recognition strategy can be improved by exploiting prosodic structural information. Although we consider evidence of the human processing of speech to be of primary importance, the development of any practical system must depend on the extraction of reliable acoustic cues. We therefore consider the exploitation of prosodic cues within a practical machine recognition framework. The development of a prosodically guided approach to continuous speech recognition is discussed, in which the prosodic structure is used to guide the segmental parsing strategy. We outline the architecture of the system and describe the implementation of the recognition strategy.

THE IMPORTANCE OF PROSODIC ORGANISATION

In many segmentally based continuous speech recognition systems a primary stage in the recognition process is the transformation of the acoustic input into a phonemic string or lattice (Hoequist and Nolan 1988, Harrington et al. 1987). Lexical entries, represented as citation form phonemic strings, are matched to the input stream in a left to right fashion, and word hypotheses formed. However, due to co-articulation and speech rate effects, sentence embedded words can differ greatly from their citation forms. Such decoding methods therefore employ many context dependent rules to provide a mapping between the ideal and the realised word forms. Unfortunately, as there is no representation of structure above the segmental level, the degradation effects must be assumed to be constant across the utterance, and this enforces an exhaustive serial recognition strategy. By allowing for all phonetically possible mappings at every contextually valid phonemic position, the approach results in a massive overgeneration of word hypotheses.

A perceptual model of human speech processing by Grosjean and Gee (1987), based upon psycholinguistic evidence, proposes a more structured approach for the perceptual process than the simple serial parsing strategy outlined above. A particular emphasis of the model is the clear processing distinction between prosodically strong and prosodically weak syllables, with strong syllables being considered to be important in conveying vital structural information to the listener. This notion of prosodic strength cannot be clearly defined in absolute terms, as the measure is a relative notion related not only to the strength of other neighbouring syllables but is also affected by rhythmic expectancy constraints imposed by the listener. However, prosodically strong syllables are generally considered to be those containing a full vowel which, due to lexical or metrical stress placement, may be realised as salient within the speech stream. Weak syllables contain vowels which are reduced in duration or distance from their normal full vowel equivalent, or are realised as the central schwa position; such syllables cannot receive stress. This distinction is reflected in the categorisation of function and content words: the, generally monosyllabic, function words (determinants, conjunctions, prepositions for example) being realised prosodically weak, but the greater information carrying content words (such as nouns, verbs, adjectives) generally being realised as, or containing, prosodically strong syllables.

Grosjean and Gee's model makes this categorisation explicit. A dual processing strategy is proposed in which strong syllables are used to initiate lexical access, whereas weak syllables are simply identified by means of a more direct pattern matching technique. In agreement with this model, Bond and Garnes provide evidence, (Garnes and Bond 1975, Bond and Garnes 1980) that prosodic strength plays a major role in invoking lexical access, with speech error data suggesting that prosodically weak syllables do not act as perceptually important units of recognition. Stressed vowels are very rarely reported incorrectly whereas the misperception of spurious syllable insertion, apparently due to expectation of weak syllable deletion, is common. Such results indicate that the low level perceptual processes have greater organisation than the simple fitting of lexical items onto the linear sequence of phonetic elements, effectively arguing against the exhaustive type of parsing strategy originally outlined above.

PROSODIC RESTRICTIONS TO PARSING

It is interesting to consider why prosodic strength is used as a cue to the importance of information within the stream. Why is it the case that those syllables which have greater prosodic strength are used for lexical access? Waibel (1986) has provided evidence towards this, in showing that prosodic strength can identify regions which have greater lexical correspondence and are therefore, in this sense, richer in information. The information content of the segmental stream is therefore not constant but varies in relation to the lexical discrimination provided by segmental context (1).

Waibel's results confirm Grosjean and Gee's model: successful parsing is achieved by invoking lexical access on the prosodically strong syllables, as it is these syllables which have a greater probability of lexical correspondence. There is however a further reason why strong syllables are considered important for speech perception, in that they are important cues to word boundary placement. Cutler and Norris (1988), and Butterfield and Cutler (1988) have indicated how strong syllables are hypothesised as, or are used to locate, potential word onsets in the human perceptual process. In addition, Cutler and Carter (1987) have shown that the statistics of spontaneous English conversation reflect this fact, ninety percent of content words having a strong syllable onset. Such data is thus valuable for the hypothesis of potential word onsets within a machine recognition system.

In considering the evidence detailed above with reference to the development of a practical machine recognition system we see that the segmental parsing strategy could be greatly improved by making the strong/weak, function/content word distinctions explicit. By classifying phonemic segments in terms of their overlying syllabic structure we can use prosodic strength to provide a measure of phonological importance, and thereby constrain the generation of word hypotheses.

PRELIMINARY RESULTS AND PARSING PROPOSAL

We have investigated and confirmed Waibel's claims of prosodic structure by an informal analysis of the segmental structure within the Cambridge "Hotel" database (2), a database of hand labelled continuous speech data. In the analysis, we examined the lexical correspondence of 2700 syllables comprising 12 repetitions of 25 phonemically balanced sentences spoken by 4 English male speakers. The prosodic relationships across the utterances, were classified as strong (approaching, or being realised as stressed), weak (approaching, or being realised as reduced), or indeterminate. These relationships were compiled by correlating the classifications of five listeners. Having excluded common intra-speaker differences due to accent, and considering the vowel reduction in weak syllables to be valid, we measured the segmental correspondence between the hand labels and the fixed lexical entry labels. We found a 13.9% error in lexical correspondence for weak positions compared to a 2.7% error for prosodically strong syllables. However, it was the nature of such errors which also proved to be of interest.

The most noticeable error for the prosodically strong syllable positions were the deletion of low energy consonants within the syllable coda, particularly weak plosives. Error within the syllabic onset of a prosodically strong syllable was particularly rare, no doubt due to the greater effort with which such segments are produced. However, for weak syllables, errors were noticed in the form of vowel deletion and consonant modification, or deletion, in both syllable onset and syllable rhyme positions. It is interesting to note that Nooteboom (1981) indicated the greater importance of the segmental structure of word onsets above word endings for lexical invocation. Considering this data in the light of the prosodic structure of word onsets detailed above, we see that both sets of data complement each other well. Furthermore, considering that the lexical correspondence data shows a tendency for strong syllable codas rather than onsets to have segmental error, such results imply that it is the onsets of prosodically strong syllables which are the major contributing factors for lexical invocation, regardless of position within the word. Obviously, a more rigorous analysis of the database is required, ensuring the validity of the hand labelled data and prosodic decisions, before any such claims could be made. Nevertheless, these regions should be the areas of focus for segmental parsing, the remainder of the parse being constrainted by the lexical and prosodic restrictions provided by these positions.

When we draw together the evidence from above, we see that for a machine recognition system we could improve the segmental parsing strategy by first employing the segmental constraints of the prosodically strong syllables, and then invoking lexical access upon these regions. Using this segmental context we can then hypothesise all contextually valid words. Furthermore, due to the syllable being prosodically strong, we can place greater weight on those words in which the syllable appears as a word onset. In matching to the lexicon, phonological rules can then be selectively applied to allow for assimilation and speech rate degradation effects mainly within the rhyme of the strong syllable, and thus generate further word hypotheses. For multisyllabic words, constraints can be applied for the other syllable positions: further phonological rules allowing reduction or deletion of the syllabic nucleus, and phonologically allowed modification, or deletion, of consonants at weak syllable positions. By restricting the otherwise exhaustive application of phonological rules in this way, and making the coarseness of the match dependent upon prosodic structure, we aim to reduce the overgeneration of word hypotheses. In this sense, the parsing strategy is seen to more closely resemble human processing, as many parses, such as those comprising strings of weak function words only (a substantial problem in linear segmental parsing), can be prevented. By exploiting the prosodic structure imposed upon continuous speech, we aim to improve upon current systems which appear to take no account of the localisation of phonemic degradation effects. In this respect the recognition of the important differences between content and function words, and the non-reliance on a strictly left to right, segment by segment, parsing method appears to provide a more satisfactory continuous speech recognition strategy.

In considering the practical implementation of a system based upon the above observations, there are two design issues: the extraction of the acoustic cues to allow such an approach, and the organisation of an architecture to accommodate the flexible parsing strategy and lexical access methods proposed. We therefore turn our attention to the practical implementation issues which have been raised during the initial development of the system.

PRACTICAL IMPLEMENTATION CONSIDERATIONS

One of the major failings of prosodic theory has been the vagueness of terminology when referenced to the acoustic domain. Syllable structure and levels of stress are in many cases freely referred to with an assumed implicity of definition. For example, although models of speech perception such as that of Grosjean and Gee refer to syllabic structure, the placement of syllable boundaries within a word is in many cases indeterminate. This obviously presents problems for any practical implementation strategy. In many cases phonological constraints can be used to allow syllable boundary location. However, as noted by Lass (1984), where there is no strong evidence for assigning a segment to one or other of its flanking syllables then there are two possible approaches: either we make some form of arbitrary theory-based decision, such as using the maximal onset principle, or we recognise the arbitrariness and ambiguity of this decision and allow ambisyllabicity. In allowing a segment to be ambisyllabic the segment is effectively shared by both syllables, and in such cases neighbouring syllables are therefore considered to overlap. Obviously in terms of the acoustic input stream, upon which the automatic recognition must be based, a decision to define some arbitrary theoretical syllabic unit will have no effect on the appearance of such units within the input stream and hence must fail. However, an acceptance of overlapping syllabic structure immediately implies that hard and fast syllable boundaries will not exist in many cases. In this system, we recognise the validity of ambisyllabicity and therefore accept that we cannot simply segment the acoustic input stream into a concatenated sequence of syllabic units.

It therefore appears that, although we need some form of syllabic chunking such units cannot always be reliably extracted from the acoustic domain. Our approach therefore does not attempt to extract syllabic boundaries from the prosodic cues, but rather relies on the phonological constraints of the input segmental stream to chunk the input into *syllable-like* (termed syllunit) strings, which are allowed to overlap. The maximal extent of any overlap is simply restricted by phonological constraints of legal syllunit onsets and endings. In this way, we can structure lexical entries in terms of such units. Obviously one problem with this approach is that a legal syllunit may span a word boundary. Furthermore, any error in the segmental input stream (3) can invalidate syllunit chunking. We overcome these problems by employing a centre-out parsing strategy from each syllunit nucleus, and also include within the lexical representation a broad phonetic class relationship allowing instant reference to other lexical entries having similar syllunit detail. The exact nature of such lexical representation and parsing will be outlined in the next section.

To locate the syllabic nuclei within the input stream, we use a detector based upon an error corrected peak-to-peak amplitude measure. Developed from a broad class segmentation algorithm originally proposed by Sergeant and Fu (1976), this algorithm gives an *false-alarm* error rate of 14% (mainly occurring in the nasal of vowel-nasal junctions within strong syllables) but a *missed* error rate of only 2.4% (occasionally missing the reduced vowel of a weak syllable). It may at first appear strange to detect nuclei in this way when the phonemic labelled input is available. However, although we can use such labels to correct for false-alarm errors, we require the prosodically based nucleus measure to locate syllunits which cannot be easily extracted from the segmental stream (such as syllabic consonants - i.e. the second syllable of *button*). Accurate location of syllabic nuclei is also required for strong/weak stress level extraction (4).

SYSTEM DESCRIPTION

To allow the implementation of the parsing strategy detailed above, we see that the lexicon must consist of more than simply a list of words represented as citation form phonemic strings. We must allow for hierarchical relationships to be made explicit between the word, syllunit, and segment levels, so that in accessing a string of segments which correspond to a legal syllunit structure we can easily access all words of which that syllunit forms a part.

Also to allow for the possibility of error in the segmental input stream, we wish to be able to access related word or syllunit sections which have the same broad phonetic structure. Finally the lexicon must also allow the prosodic structure of words to be represented along with some description of word class, i.e. whether the word generally acts as a function (generally prosodically weak) or content (prosodically strong) word.

Such a representation is naturally suited to an object oriented framework (Jackson 1986). Classes of interacting unit at phoneme, syllunit, and word levels can be naturally represented as related objects, and yet simply implemented by specifying relationships at a generic level. The representation of a small section of the lexicon, for the content words *solicitor* and *listen*, is shown in Figure 1. Note that objects such as SVL represent broad class objects, Sibilant-Vowel-Liquid for example.

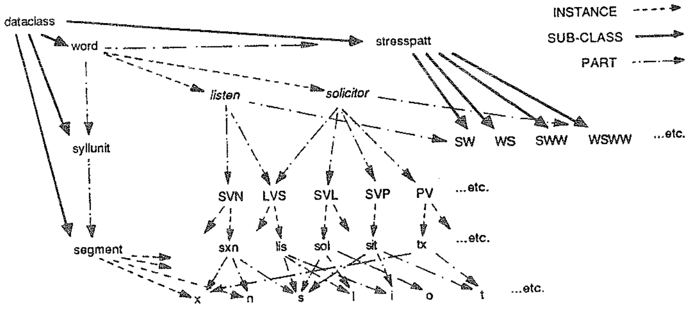

Figure 1. Representation of hierarchical lexical structure.

By representing the lexicon in terms of this large hierarchical structure we are therefore able to access any word or syllunit via segmental, prosodic, or broad class constraints. Thus any syllunit object can be investigated to access those words of which it forms a part. Similarly, once a word is accessed, the segmental formation of any neighbouring syllunit objects within the word can be immediately accessed, via the part-whole inheritance path.

Parsing proceeds by successively forming valid syllunit structures, centred on each prosodically strong nucleus. The maximal phonological unit is first assumed, and we access the lexicon for words containing that unit. The constraints generated, due to invoked content word hypotheses, are then applied to the neighbouring syllunit locations. Parsing continues by cyclicly contracting the phonological syllunit about the nucleus, and attempting further lexical access. A final phase of the parse then attempts to satisfy gaps in the input stream with reduced function word forms. In a complete speech recognition system, this final phase could be heavily constrained by syntactic rather than phonological structure.

The system is structured as a standard production system, but with knowledge sources and rules being implemented as generic classes of object. To allow implementation of such a flexible parsing strategy, the system is agenda based, and operates under the control of a generic scheduler object which supervises agenda and working memory entries, and cycles the parsing strategy. Within the limited space of this paper it is difficult to fully detail the system operation. A technical report, giving further detail of the system operation, will however be made available in mid 1989.

Unfortunately, as we are at present using a very limited lexicon of just over one hundred words, current results cannot give any true measure of the benefits of the prosodically guided parsing strategy due to the limited confusion between lexical entries. However, initial performance of the system does appear to justify this approach, and we are planning to implement a morphemically structured lexicon of over three thousand entries in the near future, and further develop the system.

NOTES

(1). For example segmental streams which correspond to morphemic stems could be considered to contain greater information than those corresponding to morphemic affixes, as identification of the former would lead to a smaller class of lexical hypotheses.

(2). Funding for the establishment of this database was provided under the STL Alvey project (MMI 069) : *Automatic speech recognition using a speech knowledge base and advanced pattern matching algorithms*, which we gratefully acknowledge.

(3). In the current system, phonemically labelled input is simulated using data from the Hotel database. The database labels can thus be considered to simulate the output of an ideal phonemic labeller. As the database has resulted from the transcription of fluent continuous speech it therefore represents common phonological degradation effects accurately, rather than being superficially modified in any way.

(4). Levels of stress are also simulated in the current system, using the the human judgements detailed in the lexical correspondence tests. However, we are intending to develop an extraction algorithm in the near future by extending Waibel's Bayesian classifier approach to measure prosodic strength in relative rather than absolute terms.

(†). Funding for the first author of this paper is provided by the Science and Engineering Research Council in the form of a Research Studentship, which we gratefully acknowledge.

REFERENCES

Bond Z. S., Garnes S. (1980) *Misperceptions of Fluent Speech*, In Perception and Production of Fluent Speech, ed. Cole et. al., Hillside, New Jersey.

Butterfield S., Cutler A. (1988) *Segmentation Errors by Human Listeners*, Proc. 7th FASE Symposium, Edinburgh.

Cutler A., Carter D. (1987) *The Prosodic Structure of Initial Syllables in English*, Proc. European Conference on Speech Tech., Edinburgh.

Cutler A., Norris D. (1988) *The Role of Strong Syllables in Segmentation for Lexical Access*, Journal of Experimental Psychology: Human Perception and Performance, 14.

Freeman F. J. (1982) *Prosody in Perception, Production and Pathologies*, in Speech Language and Hearing, ed. Lass et. al. Vol 2.

Garnes S., Bond Z. S. (1975) *Slips of the Ear: Errors in Casual Speech*, In Papers of the eleventh regional meeting of the Chicago Ling. Soc., ed. Grossman et. al., Chicago.

Grosjean F., Gee J. P. (1987) *Prosodic Structure and Spoken Word Recognition*, in Spoken Word Recognition, ed. Frauenfelder and Tyler, Ch. 6.

Harrington J. et. al. (1987) *The Application of Phoneme Sequence Constraints to Word Boundary Identification in Automatic, Continuous Speech Recognition*, Proc. European Conference on Speech Tech., Edinburgh.

Hoequist C. Jr., Nolan F.J. (1988) *An Application of Phonological Knowledge in Automatic Speech Recognition*, Submitted to Computer Speech and Language. Vol 3.

Jackson P. (1986) *The SLOOP Manual*, D.A.I. Teaching Paper No. 2, Dept. of Artificial Intelligence, University of Edinburgh.

Lass R. (1984) *Phonology*, Cambridge University Press.

Nooteboom S. G. (1981) *Lexical Retrieval from Fragments of Spoken Words*, Journal of Phonetics, Vol 9.

Sergeant D. C., Fu K. S. (1976) *Computer Algorithms for the Extraction of Stress Contours from Continuous Speech*, Report TR-EE 75-44, Purdue University, West Lafayette.

Vaissiere J. (1985) *Speech Recognition: A Tutorial*, in Computer Speech Processing, ed. Fallside and Woods.

Waibel A. (1986) *Prosody and Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University.