TOWARD A COMPUTATIONAL THEORY OF SPEECH PERCEPTION

Jonathan Allen
Research Laboratory of Electronics & Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139

ABSTRACT

In recent years,a great deal of evidence has been collected which gives substantially increased insight into the nature of human speech perception. It is the author's belief that such data can be effectively used to infer much of the structure of a practical speech recognition system. This paper details a new view of the role of structural constraints within the several structural domains (e.g. articulation, phonetics, phonology, syntax, semantics) that must be utilized to infer the desired percept.

Each of the structural domains mentioned above has a substantial "internal theory" describing the constraints within that domain, but there are also many interactions between structural domains which must be considered. Thus words like "incline" and "survey" shift stress with syntactic role, and there is a pragmatic bias for the ambiguous sentence "John called the boy who has smashed his car up." to be interpreted under a strategy that reflects a tendency for local completion of syntactic structures. It is clear, then, that while analysis within a structural domain (e.g. syntactic parsing) can be performed up to a point,interaction with other domains and integration of constraint strengths across these domains is needed for correct perception. The various constraints have differing and changing strengths at different points in an utterance, so that no fixed metric can be used to determine their contribution to the well-formedness of the utterance.

At the segmental level, many diverse cues for segmental features have been found. As many as 16 cues mark the voicing distinction, for example. We may think of each of these cues as also representing a constraint, and the strength of the constraint varies with the context. For example, stop closure duration must be interpreted in the context of the local rate of speech, and a given value of closure duration can signify either a voiced or an unvoiced stop depending on the surrounding vowel durations. Thus several cues must be integrated to obtain the perceived segmental feature, and the weights assigned to each cue vary with the local context.

From the preceding examples, it is seen that in order to model human speech perception, it is necessary to dynamically integrate a wide variety of constraints. The evidence argues strongly for an active focussed search, whereby the perceptual mechanism knows, as the utterance unfolds, where the strongest constraint strengths are, and uses this reliable information, while ignoring "cues" that are unreliable or non-determining in the immediate context. For example, shadowing experiments have shown that listeners (performing the shadowing task) can restore disrupted words to their original form by using semantic and syntactic context, thus demonstrating the integration process. Furthermore, techniques are now available for analytically finding that information in an input stimulus which can maximally discriminate between two candidate prototypes, so that the perceptual control structure can focus only on such information to make a choice between the candidates. In this paper, we develop a theory for speech recognition which contains the required dynamic integration capability coupled with the ability to focus on a restricted set of cues which has been contextually selected.

The model of speech recognition which we have developed requires, of course, an initial low-level analysis of the speech waveform to get started. We argue from the recent psycholinguistic literature that stressed syllables provide the required entry points. Stressed syllable peaks can be readily located, and use of the phonotactics of segmental distribution within syllables, together with the relatively clear articulation of syllable-initial consonants, allows us to formulate a robust procedure for determining initial segmental "islands", around which further analysis can proceed. In fact, there is evidence to indicate that the human lexicon is organized and accessed via these stressed syllables. The restriction of the original analysis to these stressed syllables can be regarded as another form of focussed search, which in turn leads to additional searches dictated by the relative constraint strengths of the various domains contributing to the percept. We argue that these views are not only consonant with the current knowledge of human speech perception, but form the proper basis for the design of high-performance speech recognition systems.