MODELS OF PHONETIC RECOGNITION III: THE ROLE OF ANALYSIS BY SYNTHESIS IN PHONETIC RECOGNITION

Victor W. Zue

Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract This paper proposes a recognition model that attempts to deal with variabilities found in the acoustic signal. The input speech signal is first transformed into a representation that takes into account known properties of the human auditory system. From various stages of this transformation, acoustic parameters are extracted and used to classify the utterance into *broad* phonetic categories. The outcome of this analysis is used for lexical access. The constraints imposed by the language on possible sound patterns should significantly reduce the number of word candidates. Finally, detailed acoustic cues will be utilized to select the correct word from the small set of candidate words.

Introduction

The task of phonetic recognition can be stated broadly as the determination of the transformation of the *continuous* acoustic signal into a *discrete* representation that can then be used for lexical access. In presenting my arguments, I will assume that words in the lexicon are represented by a set of phonological units. While the precise nature of these units, be they metrical feet, syllables, phonemes, or distinctive feature bundles, is not important for the present discussion, for the sake of consistency I will assume that words are expressed as strings of phonemes.

My proposed model of phonetic recognition makes use of broad phonetic analysis and language-specific constraints to reduce the number of lexical hypotheses, and to establish the context for further, detailed phonetic analysis. This is the third of a set of three papers from the MIT Speech Communication Group, expressing somewhat opposing views on the topic. Upon closer examination, however, there may not be as many differences as there are similarities. Like Klatt (these proceedings), I believe that the signal must be transformed into an acoustic, segmental description. However, I do not share his view regarding the feasibility of lexical access from short-time spectra, nor the use of a set of uniform distance metrics to measure phonetic similarities. Like Stevens (these proceedings), I believe in a representation based on distinctive features. However, I am increasingly frustrated by our inability to find invariance of these features in the acoustic domain, and thus I question the hypothesis that such invariance in fact exists.

Why Is Phonetic Recognition Difficult?

Phonetic recognition is difficult chiefly because the process of phonetic encoding in the acoustic signal is highly variable. Specifically, the acoustic realizations of a given phoneme can vary greatly as a function of context (Zue, 1985). On the one hand, different acoustic cues can signify the same underlying phonological representation. For example, the acoustic realization of the phoneme /t/ is drastically different in words such as "tea," "tree," "steep," "button," and "butter." On the other hand, the same acoustic cue can signify influences from different levels of the linguistic representation. For example, duration of a phoneme can be influenced by factors ranging from semantic novelty and syntactic structure to phonetic context and physiological constraints (Klatt, 1976). In order to perform phonetic decoding, a computer must extract and selectively attend to many acoustic cues, interpret their significance in light of other evidence, and combine the inferences to reach a decision. This is an immensely difficult task, given the incomplete state of our knowledge about the important acoustic cues and the ways they should be combined.

In addition to contextual variations, there are several other sources of variability that can affect the acoustic realization of utterances (Klatt, 1986). First, *acoustic variations* can arise from changes in the environment or in the position and characteristics of the transducer. Second, *within-speaker variations* can result from changes in the speaker's physiological or psychological state, speaking rate, or voice quality. Third, differences in sociolinguistic background, dialect, and vocal tract size and shape can contribute to *across-speaker variations*. Some of these variations may have little effect on phonetic distinctiveness, whereas others will have dire consequences. Successful phonetic recognition crucially depends on our ability to deal with all these sources of variability. Not only must we extract and utilize information from phonetic variations during recognition, we must also learn to disregard or deemphasize acoustic variations that are irrelevant.

Utilising Constraints

The contextual variations observed in the speech signal can often be attributed to constraints imposed by the human articulatory mechanisms. For example, the motion of the formant frequencies during the production of the diphthong /a^y/ directly reflects the movement of the tongue from a low posterior position to a high anterior position. However, superimposed on such articulatory constraints is the knowledge possessed by a native speaker that certain gestures need not be as precise as others. In American English, for example, a speaker can choose to nasalize vowels at will, since the degree of nasality does not affect a phonetic decision. Similarly, a native speaker can produce a front, rounded vowel in place of a back, rounded vowel (as in the word sequence "two two") simply because the [+back] is a redundant feature for rounded vowels in American English.

Examples of such language-specific constraints are easy to find. The so-called *phonotactic* constraints govern the permissible phoneme combinations. There are also the *prosodic* constraints, limiting the possible stress patterns for a word. Knowledge about these constraints is presumably very useful in speech communication, since it enables native speakers to fill in phonetic details that are otherwise unavailable or distorted. Evidence of the usefulness of such language-specific knowledge can be gleaned from experiments in which phoneticians were asked to transcribe utterances (Shockey and Reddy, 1975). The transcription error was typically high when the utterance was from a language unknown to the transcriber, suggesting that "knowing what to expect" is important for phonetic decoding.

Large dictionaries have been used in several recent investigations into the magnitude of phonotactic and prosodic constraints for American English and other languages (Shipman and Zue, 1982; Huttenlocher and Zue, 1984; Carlson et al., 1985). All of these studies found that a broad phonetic representation roughly corresponding to manner of articulation of phonemes can often map words into equivalence classes with extremely sparse membership. In American English, for example, the expected value of the class size based on a six-category classification scheme was found to be 34, a reduction of more than two orders of magnitude from the size of the original lexicon. Results such as these suggest that a complete and detailed phonetic analysis of the speech signal not only is undesirable but may indeed be unnecessary. Broad phonetic analysis by its nature focuses on acoustic cues that are more invariant against contextual influences. That such a

representation is also able to capture important phonological constraints imposed by the language suggests that large-scale lexical candidate reduction may be possible. Furthermore, because the exact phonetic context is specified by the candidate words, detailed phonetic knowledge can be used with greater confidence. If "tree" is a candidate word, then the verification process can use the predictive knowledge of the retroflexed context, as specified by the following /r/. The recognition algorithm will then be able to focus its attention on the detection of the retroflexed /t/ rather than a generic /t/.

A Phonetic Recognition Model

Figure 1 shows a possible recognition model incorporating some of the previously discussed ways of dealing with variability. The input speech signal is first transformed into a representation that takes into account known properties of the human auditory system, such as critical-band frequency analysis, dynamic range compression, temporal and frequency masking, adaptation and onset enhancement, and synchrony processing (see, for example, Seneff, 1985). From various stages of this transformation, acoustic parameters are extracted and used to classify the utterance into broad phonetic categories. The coarse classification also includes prosodic analysis that identifies regions where the speech signal is likely to be more robust. The outcomes of these analyses are used for lexical access. The constraints imposed by the language on possible sound patterns should significantly reduce the number of word candidates. Once the phonetic context has been established, detailed acoustic cues can then be used to select the correct answer from the small set of candidate words.

Note that the proposed recognition model is essentially a hypothesis-test, or analysis-by-synthesis, model. It has been proposed in the past for speech analysis (Bell et al., 1961) as well as for speech perception (Stevens and House, 1970). The

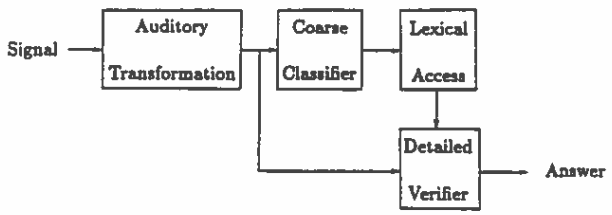

Figure 1: A Speech Recognition Model

A proposed speech recognition model that attempts to incorporate features for dealing with variabilities.

success of such a model relies heavily on the assumption that the number and the dimensionality of the hypotheses remain small. In our case, this is achieved through large-scale hypothesis pruning utilizing a proper set of constraints. Once the number of hypotheses becomes manageable, attention can be directed toward detailed acoustic cues that will enable us to make fine phonetic distinctions. The model is also computationally efficient since detailed acoustic cues are computed only when necessary. During verification, the acoustic cues can be determined in a prioritized manner as well. The computational savings, however, should be considered a side benefit; the primary appeal of the model stems from its ability to deal with variability. The coarse analysis is desirable because the resulting representation is relatively invariant across contexts and yet implicitly captures lexical and phonotactic constraints. Since detailed phonetic recognition is often error-prone, deferring this process will minimize error propagation.

To successfully implement such a model, mechanisms must be provided to insure that correct word candidates are not accidentally pruned and irretrievably lost. Errors of this sort occur for two reasons: either the coarse classifier makes a mistake or the lexicon does not anticipate a particular phonetic realization for the word by the speaker. This problem can be alleviated by permitting the lexical access procedure to accept reasonable insertions, deletions, and substitutions. If the errors are indeed reasonable, the correct word candidates should have better scores than the incorrect ones.

While the discussion leading to this model has focused on isolated words, the model can, in principle, deal with continuous speech as well. Instead of working with a set of word candidates, the verifier would deal with a *lattice* of word candidates. Provisions would then be made to determine and compare the relative goodness of words and word strings, subject to phonological, syntactic, and semantic constraints. Recent lexical studies using larger linguistic units such as syllables and metrical feet (Huttenlocher and Withgott, personal communication) show that these units exhibit constraints of similar magnitude. Using these large units may prove to be a more elegant way of accommodating continuous speech.

[Research Supported by DARPA under contract N00014-82-K-0727, monitored throught the Office of Naval Research.]

References

Bell, C. G., Fujisaki, H., Heinz, J. M., and Stevens, K. N. (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725–1736.

Carlson, R., Elenius, K., Granstrom, B., and Hunnicutt, S. (1985), "Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages," *Speech Transmission Laboratory Quarterly Progress Report*, STL-QPSR 1-2.

Huttenlocher, D. P., and Zue, V. W. (1984), "A Model of Lexical Access Based on Partial Phonetic Information," *Proc. ICASSP-84*, pp. 26.4.1–26.4.4.

Klatt, D. H. (1976), "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," *J. Acoust. Soc. Am.*, vol. 59, no. 5, pp. 1208–1221.

Klatt, D. H. (1986), "The Problem of Variability in Speech Recognition and in Models of Speech Perception," in *Variability and Invariance in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds., Hillsdale, NJ: Lawrence Erlbaum Assoc., pp. 300–319.

Seneff, S. (1985), "Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model," Ph.D. Thesis, Massachusetts Institute of Technology.

Shipman, D. W., and Zue, V. W. (1982), "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proc. ICASSP-82*, pp. 546–549.

Shockey, L., and Reddy, D. R. (1975), "Quantitative Analysis of Speech Perception," in *Proceedings of the Stockholm Speech Communication Seminar*, G. Fant, Ed., New York: John Wiley and Sons.

Stevens, K. N., and House, A. S. (1970), "Speech Perception," in *Foundations of Modern Auditory Theory*, J. Tobias and E. Schuber, Eds., New York: Academic Press.

Zue, V. W. (1985), "The Use of Speech Knowledge in Automatic Speech Recognition," *Proceedings IEEE*, vol. 73, no. 11, pp. 1602–1615.