PHONOTACTIC AND LEXICAL CONSTRAINTS IN SPEECH RECOGNITION

Daniel P. Huttenlocher and Victor W. Zue
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

We demonstrate a method for partitioning a large lexicon into small equivalence classes, based on sequential phonetic and prosodic constraints. The representation is attractive for speech recognition systems because it allows all but a small number of word candidates to be excluded, using only gross phonetic and prosodic information. The approach is a robust one in that the representation is relatively insensitive to phonetic variability and recognition error.

INTRODUCTION

Speech is the output of a highly constrained system. While it has long been recognized that there are multiple sources of constraint on speech production and recognition, natural language research has tended to focus on the syntactic, semantic, and discourse levels of processing. We believe that constraints at the phonological and lexical levels, although less well understood, are as important in recognition as higher level constraints. For a given language, the speech signal is produced with a limited inventory of possible sounds, and these sounds can only be combined in certain ways to form meaningful words. Knowledge about such constraints is implicitly possessed by native speakers of a given language. For example, an English speaker knows that "vnuk" is not an English word because it violates the *phonotactic* rules governing the allowable sound sequences of the language. He or she also knows that if an English word starts with three consonants, then the first consonant must be an /s/, and the second consonant must be either /p/, /t/, or /k/. On the other hand "smeck" is a permissible sequence of sounds in English, but is not a word because it is not in the lexicon. Such phonotactic and lexical knowledge is presumably important in speech recognition, particularly when the acoustic cues to a speech sound are missing or distorted. Perceptual data demonstrate the importance of these lower level phonological and lexical constraints. First, people are good at recognizing isolated words, where there are no higher-level syntactic or semantic constraints [3]. Second, trained phoneticians are rather poor at phonetically transcribing speech from an unknown language, for which they do not possess the phonotactic and lexical knowledge [11].

* Research supported by the Office of Naval Research under contract N00014-82-K-0727 and by the System Development Foundation

Perceptual data demonstrate that phonotactic and lexical knowledge is useful in speech recognition, we are concerned with *how* such knowledge can be used to constrain the recognition task. In this paper we investigate some phonotactic and lexical constraints by examining certain properties of large lexicons. First we consider the effects of representing words in terms of broad phonetic classes rather than specific phones. Then we discuss how this representation handles some common problems in speech recognition such as acoustic variability, and segment deletion.

PHONOTACTIC CONSTRAINTS CAN BE EXTREMELY USEFUL IN LEXICAL ACCESS

Most of the phonological rules informally gathered by linguists and speech researchers are specified in terms of broad phonetic classes rather than specific phones. For example, the homorganic rule of nasal-stop clusters specifies that nasals and stop consonants must be produced at the same place of articulation. Thus we have words like "limp" or "can't", but not "limt" or "canp". In speech perception, there is also evidence that people use knowledge about the broad classifications of speech sounds. For example, the non-word "shpeech" is still recognizable as the word "speech", while "tpeech" is not. This is because "s" and "sh" both belong to the same class of sounds (the strong fricatives), while "t" belongs to a different class (the aspirated stops). The perceptual similarity of these broad phonetic classes has long been known [9]. These broad classes are based on the so called *manner* of articulation differences. For example, the stop consonants /p/, /t/, and /k/ are all produced in the same manner, with closure, release and aspiration. The stops differ from one another in their respective *place* of articulation, or the shape of the vocal tract and position of the articulators. Manner differences tend to have more robust and speaker-invariant acoustic cues than place differences [8]. This makes broad manner classes attractive for recognition systems. However, until quite recently little was known about the role these constraints play in recognition [1] [14]. Therefore, speech recognition and understanding systems have not made much use of this information [4] [5].

Although the importance of phonotactic constraints has long been known, the magnitude of their predictive power was not apparent until Shipman and Zue reported a set of studies recently [10]. These studies examined the phonotactic constraints of American English from the phonetic distributions in the 20,000-word Merriam Webster's Pocket Dictionary. In one

study the phones of each word were mapped into one of six broad phonetic categories: vowels, stops, nasals, liquids and glides, strong fricatives, and weak fricatives. Thus, for example, the word "speak", with a phonetic string given by /spik/, is represented as the pattern:

[strong-fricative][stop][vowel][stop]

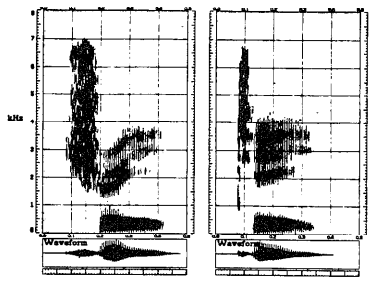
It was found that, even at this broad phonetic level, approximately 1/3 of the words in the 20,000-word lexicon can be uniquely specified. One can view the broad phonetic classifications as partitioning the lexicon into equivalence classes of words sharing the same phonetic class pattern. For example, the words "speak" and "steep" are in the same equivalence class. The average size of these equivalence classes for the 20,000-word lexicon was found to be approximately 2, and the maximum size was approximately 200. In other words, in the worst case, a broad phonetic representation of the words in a large lexicon reduces the number of possible word candidates to about 1% of the lexicon. Furthermore, over half of the lexical items belong to equivalence classes of size 5 or less. This distribution was found to be fairly stable for lexicons of about 2,000 or more words, for smaller lexicons the specific choice of words can make a large difference in the distribution.

HOW ROBUST IS A BROAD PHONETIC REPRESENTATION?

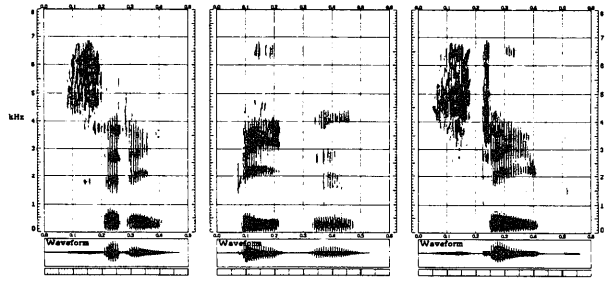
The above results demonstrate that broad phonetic classifications of words can, in principle, reduce the number of word candidates significantly. However, the acoustic realization of a phone can be highly variable, and this variability introduces a good deal of recognition ambiguity in the initial classification of the speech signal [6] [7] [12]. At one extreme, the acoustic characteristics of a phoneme can undergo simple modifications as a consequence of contextual and inter-speaker differences. Figure 1 illustrates the differences in the acoustic signal for the various *allophones* of /t/ in the words "tree", "tea", "city", and "beaten". At the other extreme, contextual effects can also produce severe modifications in which phonemes or syllables are deleted altogether. Thus, for example, the word "international" can have many different realizations, some of which are illustrated in Figure 2. Not only may phonemes be deleted, some pronunciations of a word may even have a different number of syllables than the clearly enunciated version.

In order to evaluate the viability of a broad phonetic class representation for speech recognition systems, two major problems must first be considered. The first problem is that of mis-labeling a phonetic segment, and the second problem is the deletion of a segment altogether. It is important to note that these phenomena can occur as a consequence of the high level of variability in natural speech, as well as resulting from an error by the speech recognition system. That is, not only can the recognizer make a mistake, a given speaker can utter a word with changed or deleted segments. Therefore, even a perfect recognizer would still have "errors" in its input. We address segmental variation and segmental deletion errors in the next two sections.

The scheme proposed by Shipman and Zue can handle allophonic variations, such as the different realizations of /t/.

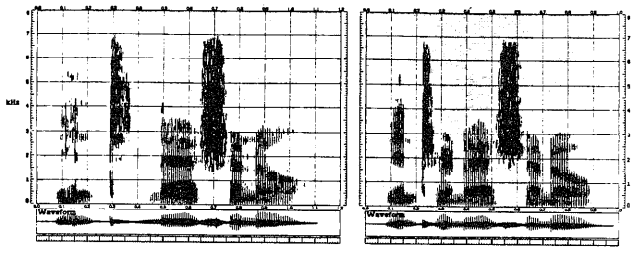


tree tea

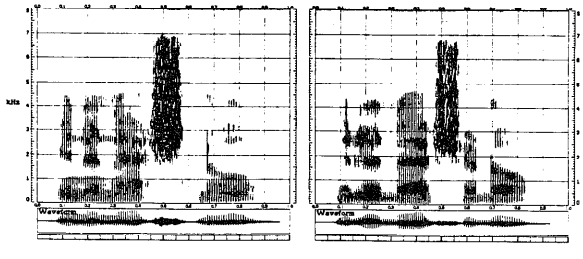


city beaten steep

Figure 1:
Spectrograms Illustrating the Acoustic Realizations of the Various Allophones of /t/



int-na-tion-al in-ter-na-tion-al



in-ner-nash-nal in-ner-na-tion-al

Figure 2:
Spectrograms Illustrating Several Possible Pronunciations for the Word "International"

This is because contextual variations tend to affect the detailed acoustic realizations of the phonetic segments, as opposed to the gross manner features used in the broad classes. When accessing the lexicon based on broad phonetic classification, detailed allophonic differences are completely disregarded. On the other hand, some uncertainties due to inter-speaker differences and recognizer errors are bound to occur. Given such uncertainties, one may ask whether the original results of Shipman and Zue still hold for lexical access.

There are a number of ways such a question can be answered. In one study we inferred the effect of these labeling ambiguities by allowing a fixed percentage of the phonetic segments in the lexicon to be unclassified while assuming that the remaining segments are classified correctly. Thus, for example, a 10% phonetic uncertainty is simulated by assuming that 90% percent of the phonetic segments in the lexicon are classified correctly. The remaining 10% of the segments can effectively be matched to any of the six phonetic categories. In order to accommodate such ambiguities, words in the lexicon must now contain not only the correct broad phonetic representation, but also those representations resulting from including unclassified segments. Admittedly our assumptions are not completely realistic, since labeling uncertainties do not occur for only a fixed percentage of the segments. Furthermore, labeling uncertainties usually arise among *subsets* of the broad categories. For example, it may be possible to confuse a strong fricative with a weak one, but not a strong fricative with a vowel. Nevertheless, we believe that such a simulation provides a glimpse of the effect of labeling uncertainties.

Table 1 compares the lexical distributions obtained from the original results of Shipman and Zue (in the first column) with those obtained by allowing 10% and 20% labeling uncertainty (in the second and third columns). The results indicate that, even allowing for a good deal of classification ambiguity, lexical constraints imposed by sequences of broad phonetic classes are still extremely powerful. In all cases, over 30% of the lexical items can be uniquely specified, and over 50% of the time the size of the equivalence class is 5 or less. On the other hand, the maximum sizes of the equivalence classes grow steadily as the amount of labeling uncertainty increases.

	Whole Word	10% Label. Errors	20% Label. Errors
% Uniquely Specified	32%	32%	32%
% In Classes of Size 5 or Less	56%	56%	55%
Max Class Size	210	278	346

Table 1: Comparison of Lexical Constraint with and without Labeling Uncertainty

PROSODIC INFORMATION CAN ALSO AID LEXICAL ACCESS

The broad phonetic class representation cannot handle segment or syllable deletions, since when a segment deletion occurs, the broad phonetic class sequence is affected. Traditionally, this problem is solved by expanding the lexicon via phonological rules, in order to include all possible pronunciations of each word [13]. We find this alternative unattractive for several reasons. For example, dictionary expansion does not capture the nature of phonetic variability. Once a given word is represented as a set of alternate pronunciations, the fact that certain segments of a word are highly variable while others are relatively invariant is completely lost. In fact, below we see that the less variable segments of a word provide more lexical constraint than those segments which are highly variable. Another problem with lexical expansion is that of assigning likelihood measures to each pronunciation. Finally, storing all alternate pronunciations is computationally expensive, since the size of the lexicon can increase substantially.

Some segments of a word are highly variable, while others are more or less invariant. Depending on the extent to which the variable segments constrain lexical access, it might be possible to represent words only in terms of their less variable parts. For instance, in American English most of the phonological rules apply to unstressed syllables. In other words, phonetic segments around unstressed syllables are more variable than those around stressed syllables. Perceptual results have also shown that the acoustic cues for phonetic segments around unstressed syllables are usually far less reliable than around stressed syllables [2]. Thus, one may ask to what extent phones in unstressed syllables are necessary for speech recognition.

In an attempt to answer this question, we compared the relative lexical constraint of phones in stressed versus unstressed syllables. In one experiment, we classified the words in the 20,000-word Webster's Pocket Dictionary either according to only the phones in stressed syllables, or according to only the phones in unstressed syllables. In the first condition, the phones in stressed syllables were mapped into their corresponding phonetic classes while the entire unstressed syllables were mapped into a "placeholder" symbol. In the second condition the opposite was done. For example, in the first condition the word "paper", with the phonemic string **/'pe^{v}-pə'/**, is represented by the pattern:

[stop][vowel][*]

where * is the unstressed syllable marker. In the second condition the word is represented by the pattern:

[*][stop][vowel]

where * is the stressed syllable marker. Note that at first glance a significant amount of information is lost by mapping an entire syllable into a placeholder symbol. Closer examination reveals, however, that the placeholder symbol retains the prosodic structure of the words. A representation which makes this more explicit combines the partial phonetic classification with syllabic stress information. Thus, in the first condition, the word "paper"

would be represented as:

[stop][vowel] + [S][U]

where [S] and [U] correspond to stressed and unstressed syllables, respectively.

The results of this experiment are given in the second two columns of Table 2. The results summarized in the table are obtained by explicitly representing the prosodic information as sequences of stressed and unstressed syllables. The results for "wildcarding" the deleted syllables are almost identical and hence are not presented here. The first column of the table gives the results for the whole word (as in [10]). The second and third columns show the cases where phonetic information is only preserved in the stressed or in the unstressed syllables. It should be noted that the results cannot be accounted for simply on the basis of the number of phones in stressed versus unstressed syllables. For the entire lexicon, there are only approximately 1.5 times as many phones in stressed than in unstressed syllables. In addition, if one considers only polysyllabic words, there are almost equal numbers of phones in stressed and unstressed syllables, yet the lexical distribution remains similar to that in Table 2.

These results demonstrate that the phonotactic information in stressed syllables provides much more lexical constraint than that in unstressed syllables. This is particularly interesting in light of the fact that the phones in stressed syllables are much less variable than those in unstressed syllables. Therefore, recognition systems should not be terribly concerned with correctly identifying the phones in unstressed syllables. Not only is the signal highly variable in these segments, making classification difficult; the segments do not constrain recognition as much as the less variable segments.

This representation is very robust with respect to segmental and syllabic deletions. Most segment deletions, as was pointed out above, occur in unstressed syllables. Since the phones in unstressed syllables are not included in the representation, their deletion or modification is ignored. Syllabic deletions occur exclusively in unstressed syllables, and usually in syllables containing just a single phone. Thus, words with a single-phone unstressed syllable can be stored according to two syllabic stress patterns. For example the word "international" would be encoded by the phones in its stressed syllables:

[vowel][nasal][nasal][vowel][strong-fricative]

with the two stress patterns [S][U][S][U]U] and [S][U][S][U] for the 5- and 4-syllable versions. The common pronunciations of "international" (e.g., those in Figure 2) are all encoded by these two representations, while unreasonable pronunciations like "interashnel" are excluded.

SUMMARY

We have demonstrated a method for encoding the words in a large lexicon according to broad phonetic characterizations. This scheme takes advantage of the fact that even at a broad level of description, the sequential constraints on allowable sound sequences are very strong. It also makes use of the fact that the phonetically variable parts of words provide much less lexical constraint than the phonetically invariant parts. The interesting properties of the representation are that it is based on relatively robust phonetic classes, it allows for phonetic variability, and it partitions the lexicon into very small equivalence classes. This makes the representation attractive for speech recognition systems [15].

	Whole Word	Stressed Syls.	Un-Stressed Syls.
% Uniquely Specified	32%	17%	8%
% In Classes of Size 5 or Less	56%	38%	19%
Average Class Size	2.3	3.8	7.7
Max Class Size	210	291	3717

Table 2: Comparison of Lexical Constraint in Stressed vs Unstressed Syllables

Using a broad phonetic representation of the lexicon is a search avoidance technique, allowing a large lexicon to be pruned to a small set of potential word candidates. An essential property of such a technique is that it retains the correct answer in the small candidate set. We have demonstrated that, for a wide variety of speech phenomena, a broad phonetic representation has this property.

REFERENCES

[1] Broad, D.J. and Shoup, J.E. (1975) "Concepts for Acoustic Phonetic Recognition" in R.D. Reddy, *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*. Academic Press, New York.

[2] Cutler, A. and Foss, D.J. (1977) "On the Role of Sentence Stress in Sentence Processing", Language and Speech, Vol. 20, 1-10.

[3] Dreher, J.J. and O'Neill, J.J. (1957) "Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases", Journal of the Acoustical Society of America, vol. 29, no. 12.

[4] Erman, L.D, Hayes-Roth, F., Lesser, V.R., and Reddy, R.D. (1980). "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty", Computing Surveys, vol. 12, no. 2, 213-253.

[5] Klatt, D.H. (1977) "Review of the ARPA Speech Understanding Project", Journal of the Acoustical Society of America, vol. 62, no. 6, 1345-1366.

[6] Klatt, D.H. (1980) "Speech Perception: A Model of Acoustic Phonetic Analysis and Lexical Access" in R. Cole, *Perception and Production of Fluent Speech*. Lawrence Erlbaum Assoc., Hillsdale, N.J.

[7] Klatt, D.H. (1983) "The Problem of Variability in Speech Recognition and in Models of Perception". Invited paper at the 10th International Congress of Phonetic Sciences.

[8] Lea, W.A. (1980) *Trends in Speech Recognition*. Prentice-Hall, N.Y.

[9] Miller, G.A. and Nicely, P.E. (1954) "An Analysis of Perceptual Confusions Among Some English Consonants", Journal of the Acoustical Society of America, vol. 27, no. 2, 338-352.

[10] Shipman, D.W. and Zue, V.W. (1982) "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems", Conference Record, IEEE International Conference on Speech Acoustics and Signal Processing, Paris, France, 546-549.

[11] Shockey, L. and Reddy, R.D. (1974) "Quantitative Analysis of Speech Perception: Results from Transcription of Connected Speech from Unfamiliar Languages" *Speech Communication Seminar*, G. Fant (Ed).

[12] Smith, A. (1977) "Word Hypothesization for Large-Vocabulary Speech Understanding Systems", Doctoral Dissertation, Carnegie-Mellon University, Department of Computer Science.

[13] Woods, W. and Zue, V.W. (1976) "Dictionary Expansion via Phonological Rules for a Speech Understanding System", Conference Record, IEEE International Conference on Speech Acoustics and Signal Processing. Phila, Pa. 561-564.

[14] Zue, V.W. (1981) "Acoustic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments", Proceedings of the 1981 NATO Advanced Summer Institute on Speech Recognition, Bonas, France.

[15] Zue, V.W. and Huttenlocher, D.P. (1983) "Computer Recognition of Isolated Words from Large Vocabularies". IEEE Computer Society Trends and Applications Conference. Washington, D.C., 121-125.