Prosodic Knowledge Sources for Word Hypothesization in a Continuous Speech Recognition System

Alex Waibel

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Previously we have reported on the extraction of prosodic cues (such as stress, pitch, duration) from continuous speech [1] and have reported on possible uses of some prosodic information (e.g., temporal cues [2]) in large vocabulary word recognition systems. In this paper we extend these previous findings to a speaker-independent continuous speech recognition system. Speaker-independent knowledge sources (KS) were implemented that attempt to hypothesize words based on only prosodic cues found in the signal. The prosodic cues exploited were temporal cues (syllable durations, ratios of unvoiced segment durations to syllable durations, voiced segment durations), intensity profiles and likelihoods of stressedness. Each KS extracts the appropriate prosodic cue and searches its knowledge base for words whose prosodic patterns satisfy the constraints found in the signal. Usign a multispeaker continuous speechdatabase for evaluation, each prosodic KS is shown to hypothesize the correct word substantially better than chance. All prosodic KSs were then combined and compared with a speaker-independent acoustic-phonetic word hypothesizer. After applying the prosodic KSs, the correct word ranked on average 25th (out of 252 words). The acoustic-phonetic KS alone yielded an average rank of 40 (out of 252) without the addition of prosodic information. After prosodic and phonetic KSs were combined the average rank was reduced to 15 out of 252. The results indicate that prosodic information indeed adds complementary information that substantially improves word hypothesization in speaker-independent continuous speech recognition systems.

1. Introduction

To this day, the prosodic cues in the speech signal, duration, rhythm, intensity, pitch, and stress, are frequently being ignored in the implementation of speech recognition systems. In systems aimed at small vocabulary sizes, most research has centered around suitable representations of spectral information and around optimal search procedures used to align the unknown pattern with reference word template. In large vocabulary continuous speech recognition systems, atomic units of speech smaller than the word are usually chosen and recognition is performed by detecting and assembling phonemic or phone like units into strings of hypothesized words. Several attempts at using prosodic cues in speech recognition systems have mostly been limited to aiding syntactic analysis by hypothesizing phrase or clause boundaries (from pitch excursions) and/or hypothesizing phonemically reliable parts of the utterance ("islands of reliability") from the amount of stress found in the signal [3]. Only a few studies have attempted to use these cues to aid in the hypothesization or verification of words in English, despite the known strong contributions of prosodic cues to human word perception (see [4, 5] for a review). For isolated large vocabulary word recognition it has been shown [2, 6] that temporal cues can indeed be used effectively to hypothesize words, even in the absence of phonetic information. Moreover, these prosodic cues are shown to be predictable such that all necessary reference information for particular word candidates could be synthesized from text [2, 5]. These results, however, were limited to speaker dependent isolated word recognition and used only the temporal information in the signal.

In this paper we expand on these encouraging findings along several dimensions. First, we explore three separate prosodic parameters. In addition to temporal cues, we will use intensity and stress patterns as descriptors of the word. Second, we will be using two continuous speech databases. The former, a training and development database, consists of 50 Harvard sentences [7] and was recorded and hand-labelled at CMU. The latter, the testing database, consists of two sets of these 50 Harvard sentences, read by different speakers at MIT. The third dimension, finally is the speaker dimension. All development and testing will be performed using multiple speakers for our results to measure *speaker independent* performance. Each ten sentences in the training and testing databases were therefore read by a different speaker.

The sections of this paper are organized according to prosodic cues. For each cue, a KS was developed that using only this cue attempts to hypothesize word candidates that are most likely to satisfy the detected prosodic pattern. We will report below the operation and performance of each of these KSs. We will then compare all prosodic KSs with each other and combine them into and statistically optimal combined prosodic KS. The performance of these prosodic KSs will then also be compared with a speaker-independent phonetic word hypothesizer developed at CMU. We will show that the performance of the prosodic KSs compares favorably with the performance of the phonetic KS and that the combination of the two results in dramatic overall improvements.

2. Prosodic Knowledge Sources

Conceptually, each KS described below consists of three major components: a prosodic parameter extraction algorithm, a knowledge base, and a matcher to search for suitable word candidates. The parameter extraction algorithm performs the appropriate measurements

This research was sponsored in part by the National Science Foundation, Grant MCS-7825824 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-78-C-1551.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

on the acoustic signal to obtain the relevant prosodic cues. The knowledge base contains for each word candidate one or more (to allow for alternates) entries. Each entry consists of parametric descriptions of the word in terms of the KS-specific prosodic cue. To allow for such a knowledge base to be expanded to larger vocabularies, it is also desirable that the prosodic representation of each word be valid across different speakers or that it can be automatically predicted from text without user training. The matcher, finally, uses the prosodic cue measured by the extraction algorithm and searches the knowledge base for similar tokens. This search is typically done by assigning a score to each word candidate based on the similarity of its prosodic pattern to the pattern found in the unknown signal. The list of word candidates is then ranked according to their scores. At the absence of begin/end points in continuous speech, this analysis was performed by each KS repeatedly for each possible word anchor point, given by each hypothesized syllable boundary. Using the hand-labelled speech databases described above, the ability of each KS to hypothesize words based only on prosodic cues was then evaluated. The evaluations reported below will show the rates at which the correct word candidate will be found among the N top ranking candidates.

2.1. Duration and Rhythm

Three measures of duration were explored in three KSs: the syllable durations in a word, the ratios of the duration of the unvoiced segments in a syllable to the syllable duration, and the duration of vocalic segments. A syllable boundary was defined to lie at the onset of a rise in vocalic energy. The syllable boundaries and the unvoiced/voiced segment boundaries needed for measurement of the relevant duration patterns were detected by a set of segmentation and syllabification algorithms described in detail elsewhere [2, 5]. Two knowledge bases were evaluated. The first used duration measurements obtained from the training database, i.e., the CMU-Harvard database. For the second, all durations were synthetically generated using a knowledge compiler developed earlier [2, 6].

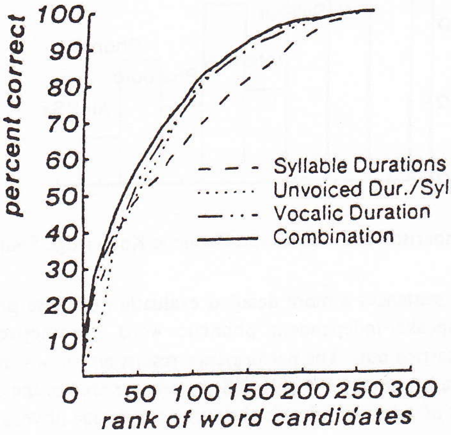

Figure 1. Percent Correct for Given Rank and for Different Durational Knowledge Sources; Testing Data.

Fig. 1 shows the results obtained by the three durational KSs. For this evaluation the testing data (100 MIT-Harvard sentences) was used. The knowledge base consisted of measured durations. All three durational measurements yield comparable performance with the syllable duration measure lagging behind somewhat. Fig. 2 shows performance results for the combination of the three durational KSs using a simple geometric mean of each KS's rank orderings. Here the effect of measured vs. synthetic knowledge base was evaluated. Also both evaluation runs were performed for both the testing and the training database. The performance degradation due to segmentation/syllabification errors can be inferred in this figure from the less than perfect performance obtained when the training data was used for both the knowledge base and as evaluation data. The inherent variability of durational cues is reflected by the additional decrement in performance when evaluation was performed using different, e.g., the testing data. Further degradation can be observed when measured durations were replaced by the synthetically generated durations. Despite these performance degrading factors, however, it is clear from this evaluation that better-than-random word hypothesization can be performed based on durational cues only.

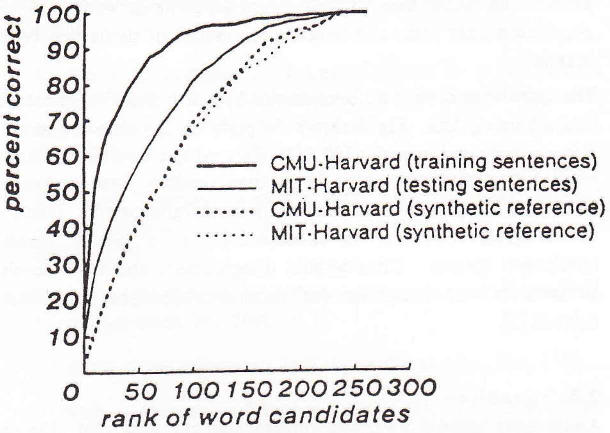

Figure 2. Percent Correct for Given Rank in Training and Testing Data; Knowledge Bases of Measured and Synthetic Durations.

2.2. Stress

Similar in spirit to the previous subsection, a KS based on stress patterns was implemented and tested. The KS uses stress probabilities

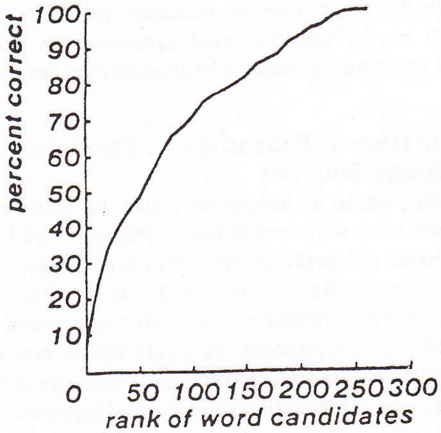

Figure 3. Percent Correct for Given Rank Using a Stress Based Knowledge Source; Testing Data.

obtained from a probabilistic stress detector [1, 5]. Thus stress *probabilities* rather than discrete stress assignments were used. This provided a finer grain and hence a continuum of similarities between tokens.

The knowledge base therefore contained stress probability as measured in the training data. Fig. 3 shows the performance obtained when this KS was evaluated over the 100 MIT-Harvard test sentences. Although word hypothesization can be better than random, these performance results are inferior to those obtained by the durational KSs. This is due to the great variability in stressedness that is indeed found in continuous speech. Considerable disagreement about the levels of stressedness was found in this data even for groups of human subjects [5].

2.3. Intensity

An intensity based KS was also implemented and evaluated. The peak-to-peak amplitude of the signal waveform was chosen as a measure of intensity. The knowledge base contained coarse amplitude patterns for the words in the vocabulary. Matching was done by measuring the similarity between the incoming patterns and the patterns in the knowledge base. Allowance was made for slight misalignments of corresponding patterns.

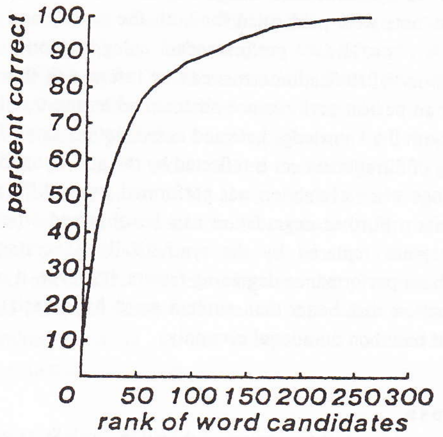

Figure 4. Percent Correct for Given Rank Using an Intensity Based Knowledge Source; Testing Data.

Fig. 4 shows the results from an evaluation run using the testing database. It can be seen that word hypothesization performance considerably better than random can be obtained from this KS.

3. Combination of Prosodic and Phonetic Knowledge Sources

In the preceding section we have demonstrated that prosodic cues can indeed be used at the word level to rank appropriate word hypotheses better than chance and speaker independently in continuous speech. In this section we would like to combine and evaluate all prosodic KSs and compare their performance with a speaker independent phonetic word hypothesizer. Furthermore, we would like to experimentally determine whether prosodic KS do lead to complementary information, that would be useful *in addition* to a phonetic word hypothesizer.

We start with the combination of prosodic KSs. To obtain a statistically optimal combination of the all five KSs described in the previous sections, we have collected variances and covariances of the scores obtained from each KS. The resulting covariance matrix was then used to compute a Mahalanobis distance as a combined prosodic similarity measure. In this fashion the contributions from each KS were weighted according to their relative merit in the light of the performance of the other competing KSs. The resulting performance graph (using the test-database) is shown in Fig. 5. Note, that the intensity KS appears to be yielding near optimal performance.

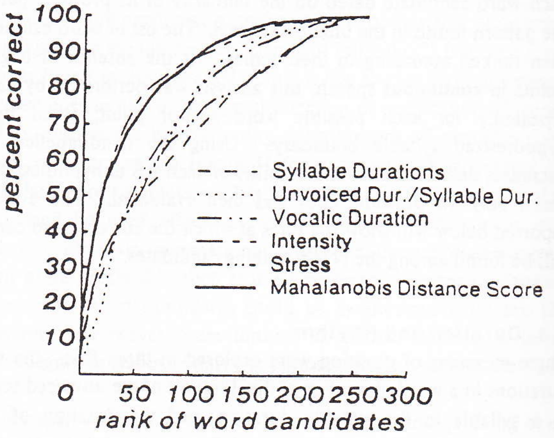

Figure 5. Combination of Prosodic Knowledge Sources

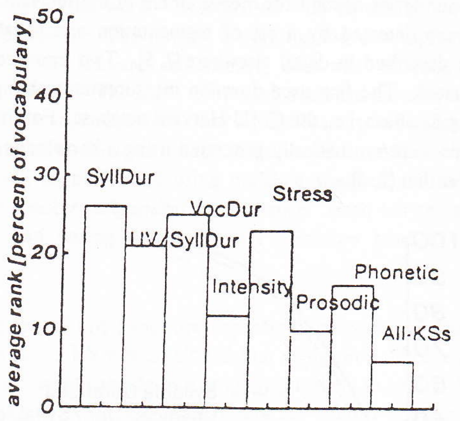

Figure 6. Comparison of Prosodic and Phonetic Knowledge Sources

Using ten test sentences a more detailed evaluation of these prosodic KSs and a speaker-independent phonetic word hypothesizer was subsequently carried out. The performance results are shown in Fig 6 in the form of a bar graph. For each KS the *average rank* of the correct word in the list of word candidate is given as a percentage of vocabulary size. Thus, for example, an average rank of 68 (for the syllable duration KS) is given as 26%, based on a vocabulary size of 252 words. From Fig. 6 we can see again that intensity patterns were the most useful prosodic cue for word identification (lowest rank). This can in part be explained by the comparatively robust prosodic parameter extraction in this case. Following the five bars representing each prosodic KS, Fig. 6 then shows the combination of all five prosodic KSs as discussed before. It is worth noting that not only was the average rank of the combined prosodic KSs better than each individual KS by itself, but that the standard deviation of the combination (not shown in this graph) was found to be considerably lower. More robust performance can therefore be expected from the exploitation of *all* cues. This combined prosodic performance measure was then compared with a speaker-independent word hypothesizer developed at CMU. It should be mentionned, that this word hypothesizer was only a preliminary version of a more advanced word hypothesizer that is currently under development. Fig. 6 shows that the rank of the combined prosodic KSs is actually lower than the phonetic word hypothesizer. Finally, combination of prosodic *and* phonetic KSs leads to substantially reduced hypothesization rank. It can be seen that adding prosodic information to the phonetic word hypothesizer reduced the average rank of the correct word hypothesis to about 1/3.

4. Conclusion

In this paper we have demonstrated that the prosodic cues of duration, intensity, and stress can be effectively used in word hypothesization. Using prosodic cues only, performance comparable or better than a speaker-independent phonetic word hypothesizer was obtained. Moreover, the combination of prosodic *and* phonetic KSs leads to dramatic improvements over phonetic word hypothesization alone. This result clearly demonstrates, that prosodic cues yield complementary information. Speech recognition systems can therefore benefit considerably from the exploitation of these cues. This paper has shown only one strategy towards achieving effective integration of prosodic analysis. Alternate strategies, such as top down verification of confusable word hypotheses are conceivable and work along these lines is in progress.

1. A. Waibel, "Recognition of Lexical Stress in a Continuous Speech Understanding System - A Pattern Recognition Approach", *ICASSP '86 Proceedings*, IEEE, 1986, pp. 2287-2290.

2. A. Waibel, "Suprasegmentals in Very Large Vocabulary Isolated Word Recognition", *ICASSP '84 Proceedings*, IEEE, 1984, pp. 26.3.1-26.3.4.

3. W.A. Lea, *Prosodic Aids to Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1980, ch. 8.

4. I. Lehiste, *Suprasegmentals*, MIT-Press, Cambridge, MA, 1970.

5. A. Waibel, *Prosody and Speech Recognition*, PhD dissertation, Computer Science Department, Carnegie Mellon University, 1986.

6. A. Waibel, *Suprasegmentals in Very Large Vocabulary Word Recognition*, In: Pattern Recognition by Humans and Machines, E.C. Schwab and H.C. Nusbaum, editors, Academic Press, Orlando, Florida 32887, 1986, ch. 5.

7. IEEE, "IEEE Recommended Practice for Speech Quality Measurements", *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-17, No. 3, September 1969, pp. 225-246, Standards Publication No. 297, available from IEEE