Suprasegmentals in Very Large Vocabulary Isolated Word Recognition

A. Waibel

Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

Prosodic information is believed to be valuable information in human speech perception, but speech recognition systems to date have largely been based on segmental spectral analysis. In this paper I describe parts of a front end to a very-large-vocabulary isolated word recognition system using prosodic information. The present front end is template independent (speaker training for large vocabulary systems (> 20,000 words) is undesirable) and makes use of robust cues in the incoming speech to obtain a presorted vocabulary of candidates. It is shown that prosodic information, e.g., the rhythmic structure of an input word, its syllabic structure, voiced/unvoiced regions in the word and the temporal distribution of back/front vowels, nasals and liquids and glides, can be used effectively to select a substantially reduced subvocabulary of candidates, before any fine phonetic analysis is attempted to recognize the word.

1. Introduction

Automatic recognition of isolated words from very large vocabularies (several thousand words) has recently received increased attention. Most current recognition systems are not easily extensible to handle vocabularies of more than a few hundred words. One problem is the great hardware cost or unacceptable slow response time when one attempts to recognize an utterance by searching a large vocabulary exhaustively using search intensive methods such as template-matching alone. Second, maintaining and collecting a database of reference word-templates becomes costly for large vocabularies and clearly impractical for many applications. Finally, a useful large vocabulary recognition system will have to be easily modifiable for the addition or subtraction of new vocabularies.

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.

Several studies have proposed useful ways to overcome some of the problems mentioned here: Preselection of possible word candidates based on simple but robust information has been shown by Kaneko and Dixon[1] to drastically reduce the remaining search space and to allow for near real-time-large-vocabulary recognition. Shipman and Zue[2] suggest that recognition of fairly simple phonetic categories such as NASAL, VOWEL, STOP, etc. can provide powerful constraints to lead to substantial vocabulary reduction. Possible alternatives to word template matching are conceivable by using demisyllable templates as the recognition unit rather than word templates[3] or creating word templates synthetically[4].

The general philosophy underlying the development of the system presented in this paper is to create experts that, based on the acoustic signal, derive robust constraints limiting the number of possible word candidates that satisfy these constraints. These experts therefore act as vocabulary filters providing an ordering of the word candidates based on their domain of expertise. The specific constraints obtained from each expert are compared with domain specific knowledge that was automatically derived and precompiled from the original orthographic spelling (text) for each vocabulary item. The proposed system therefore functions in a *template-independent* way and new vocabulary items can be entered simply in their orthographic form.

In this paper, we discuss in particular several suprasegmental vocabulary filters, briefly outline their operation and present data evaluating their current performance. It will be seen that suprasegmental cues, i.e., information that we have so far largely ignored (or warped away), can provide a complementary perspective on the speech signal that leads to considerable constraining of the possible list of word candidates. These suprasegmental cues consist of the rhythmic structure of the utterance, the temporal contribution of voiced and unvoiced regions in a syllable, and the temporal contribution of some sonorant classes, e.g., NASAL, L, R, FRONT and BACK to the duration of the syllable nucleus. Because of its importance to the

CH1945-5/84/0000-0276 $1.00 © 1984 IEEE

suprasegmental filters I start with an outline of the syllabification algorithm used. Second I describe the linear machine that serves as sonorant feature detector. Then the knowledge compiler that automatically generates the appropriate prosodic information from text will be presented. Finally, the current filters will be outlined and results of performance evaluation will be given.

2. Syllabification

Syllable boundary detection is performed in three stages. The first two use algorithms to perform general contour analysis and are applicable to any contour. They are based on techniques commonly employed in the vision and pattern-recognition literature[5, 6, 7].

In the first stage an input contour is approximated by line-segments describing only the significant events in the contour. This is done by using a recursive convex-deficiency algorithm. It starts by assuming a straight line between the begin and end points of the utterance. It finds the point P of maximum deviation of the contour from the straight line and if this deviation exceeds some stop-criterion it breaks the large line segment up into to two smaller line segments from the begin point to P and from P to the endpoint and recurses. This process continues until the deviations of the original contour from the line segment approximation can be considered insignificant. Thus the algorithm attends to increasing levels of detail from one level of recursion to the next and line segment descriptions can be extracted at varying degrees of coarseness. The algorithm is also edge preserving such that significant events in a waveform are not smeared out, but rather are preserved with their original amplitude and at their original point in time.

Following the approximation of the original contour by line segments, collections of line segments are parsed syntactically into several primitive shapes labeled Hat, Plateau, B-skirt (before a hat), A-skirt (after a hat) and Silence. These basic shapes then characterize the events in a particular contour.

Based on the contour analysis, a set of rules are applied to determine whether a boundary between subsequent events is a syllable boundary or not. At present the contours used are the smoothed peak-to-peak amplitude, the zero-crossing contour of the input signal and a sonorant energy contour. The rules take into account the basic shapes and magnitudes of the events in these contours and the possible sequences of events for a syllable to determine voiced or unvoiced portions, to find genuine syllable nuclei and finally to place the syllable boundary at a linguistically consistent point in time. The syllable boundaries are given by the onsets of syllable nuclei, which to a first approximation are known to be the points at which human listeners perceive rhythmic beats in an utterance[8]. In informal experiments with several speakers, the syllable boundary detector in its current form has been found to yield an error rate of approximately 4 - 10 %. Possible improvements might be achieved through added rules as well as alternative informative input contours.

3. The Sonorant Feature Detectors

The sonorant feature detectors described here are based on the theory of linear machines[6, 9]. A linear machine is attractive to provide sonorant categories, both because non-parametric learning can be achieved easily using error-correcting procedures (e.g. relaxation or perceptron learning) and recognition can be performed efficiently as an FIR-filtering operation. Perceptron learning has recently been successfully applied to consonant recognition[10].

For each of the categories of interest (NASAL, L, R, FRONT, BACK) relaxation[6] was used to learn a set of weights for a two-category linear discriminant function. The input features used where dependant on their relevance to classification of a particular sound. They included 54 spectral coefficients spanning an 8000 Hz spectral range, the peak-to-peak amplitude, formant frequencies as given by a formant tracker[11], and for the special case of R, the 25 spectral coefficients above F2[1] The algorithm learns the appropriate weights that best combine the given evidence (the features) on a frame by frame basis. Learning was performed using a set of 57 randomly selected hand labeled words uttered by one speaker.

In a second layer, a perceptron-based multi-category linear classifier was used to select a unique category for each frame. The input to this classifier consists of the decisions derived from the two-category response units, as described above, within a window around the current frame. Finally the output of this layered piecewise linear machine is smoothed.

4. The Knowledge Compiler

The purpose of the knowledge compiler is to generate domain specific knowledge about a particular word in a template independent way such that new vocabulary items can be easily added by simply running the compiler. An important design criterion for the compiler is to perform in the *compilation phase* most of the necessary computation needed to match the properties of an unknown speech utterance and the expected properties of a vocabulary item rather than in the recognition phase.

[1] This was found to be very useful since R's are most easily characterized by a low F3 "riding" on top of F2

Convenient representations have therefore been selected to be generated by the compiler.

The compiler consists of two major parts. The first part consists of parts of the MIT text-to-speech synthesis system[12]. It consists of reformatting of the input text, decomposition of input words into constituent morphs, phrase level parsing, letter-to-sound rules or lexicon-lookup, phonological rules and finally the generation of a phonemic representation and corresponding prosodic information (e.g. F0-target values, segmental durations, lexical stress-markers, syllable boundaries) of the input word. Parts of this system had to be changed to generate alternate pronunciations, such as for the word LETTER, where TT could be pronounced as the voiceless stop T or as a flap DX.

The second part generates additional prosodic information, improves the given information and compiles the synthetic information generated so far into a compact, useful and consistent representation. For example, syllable boundary markers are placed in the phoneme string, consistent with the syllable boundaries that are generated by the syllabification unit on incoming natural speech. Syllable durations as well as durations of voiced and unvoiced segments and of various sonorant portions of the syllable nucleus are computed for each lexical item from the segmental durations derived in the first part. The compiler presently also generates primary stress markers, expected amplitudes and formant target values. A set of rules that generate additional lexical entries for alternate syllabifications (missed boundaries, Schwa-deletions, etc.) and alternate pronunciations. The resulting compiled dictionary approximately doubles in number of entries. Further rules make adjustment to the prosodic information based on segmental context, position in the word and the number of syllables in a word. In this way, speech knowledge is incorporated in the compiler in the form of production like rules. It is our hope that the addition of further speech relevant knowledge will continue to improve recognition results.

5. Suprasegmental Vocabulary Filters - Results

Based on the information derived from the speech signal as described in the previous sections we are now ready to match them with the synthetic information given by the knowledge compiler. The following vocabulary filters have been implemented and evaluated:

- Rhythm (F1) -- various words in a large vocabulary differ by their rhythmic structure. Rhythm is therefore measured and compared with the synthetic rhythm in the database. To do so syllable durations are measured between the boundaries given by the syllable detector. The syllable durations of the natural utterance are compared with the synthetic syllable durations by normalizing for overall utterance lengths and computing a Euclidean distance.

- Voiced/Unvoiced Ratios (F2) -- The contributions of unvoiced section to overall syllable durations are measured in percent and compared with the synthetic information.

- Stressed Syllable Formant Measurement -- As an attempt to characterize the vowel nucleus of the stressed syllable in the utterance, formant frequencies have been measured[11] and compared with synthetic formant target values. This is therefore a very rudimentary segmental vocabulary filter. Stressed syllables were assumed to be the syllable with maximum peak-to-peak amplitude. No major difficulties arise in case of stress detection errors since in this case simply a potentially less reliable syllable will be considered. In order to avoid search, this filter attempts to find the steady state portions of the formant tracks and compares the measurements with the synthetic data. The major difficulties encountered with this method was to determine reliable portions of a syllable in a simple and efficient way. Nonetheless it does provide discriminatory information and was included in the evaluation.

- Nasal Contribution to Syllable Duration (F4) -- this filter and the following operates in an analogous way to filter F2. Temporal contribution of nasal portions to a the syllable nucleus are compared with synthetic data.

- R contributions to Syllable Durations (F5) -- uses sonorant feature R.

- L contributions to Syllable Durations (F6) -- uses sonorant feature L.

- FRONT contributions to Syllable Durations (F7) -- measures the contributions of front vowels to the syllable nucleus.

- BACK contributions to Syllable Durations (F8) -- measures the contributions of back vowels to the syllable nucleus.

Note that the temporal contributions of various sonorant features depend heavily on context. For example, L followed by a FRONT vowel will commonly transition through a region classified as BACK, etc. These and many other properties are included in the compiler rules that generate the expected information for a word.

A corpus of 1478 words was selected from the union of the 1000 most frequent *spoken* English words and the 1000 most frequent *written* English words[13]. The knowledge compiler thus generated prosodic information for 1478 English words. After application of the rules for alternate pronunciations and syllabification a total of 3207 vocabulary items was obtained. 57 random words from this corpus were read by one male American talker and used in the learning phase described in section 2. To evaluate the system described, 978 different words from this corpus were read by the same speaker. These 978 words were not used for the training of the classifiers nor were they considered when the compiler rules were developed. The system will therefore be tested on new English words, whose expected characteristic information was generated automatically from text. 116 of the words were found to be improperly recorded, failed to be processed accurately by the signal

processing, endpoint detection or syllabification stages and hence do not enter the results given below. The effective corpus of words tested below therefore consists of 862 utterances from a 1478 word vocabulary.

Results are given in Fig.5-1 for monosyllabic words and Fig.5-2 for polysyllabic words. For each filter (F1 through F8) we show the frequency at which the correct word is ranked among the top N candidates. The bold curves labeled C display the ranking of the correct word candidate after combination of the results from all 8 filters. Combination was performed by computing the geometric means of the individual filter rankings and reranking. Better

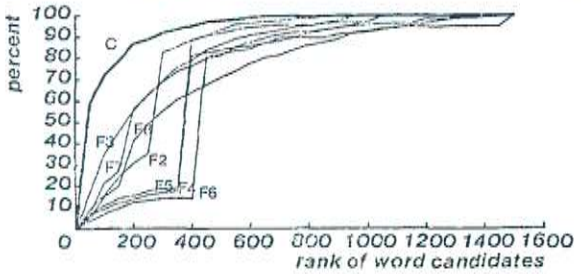

Figure 5-1: Rank of Correct Monosyllabic Word Candidates

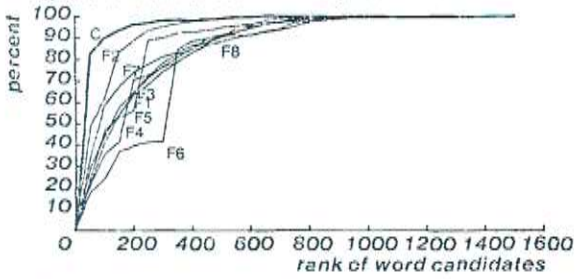

Figure 5-2: Rank of Correct Polysyllabic Word Candidates

performance was obtained for polysyllabic words as seen in Fig.5-2. This is to be expected since polysyllabic words are richer in prosodic information. In fact 27% of all polysyllabic words were uniquely identified (rank 1) by these suprasegmental filters. In contrast this is the case for only 4% of the monosyllabic words. Note also the sharp discontinuities in both figures for some of the filters. If several utterances match equally well they all received the same rank, their median rank. Hence for some filters large pools of perfectly but equally matching word candidates yield lower ranks (for example, all 769 monosyllabic dictionary entries without R-contribution are ranked 384 by filter F5 if the unknown is an utterance containing no R contribution). In summary, by applying all the constraints given by the 8 filters on the List of 3207 candidates the correct word ranks on average 91st for monosyllabic, 37th for polysyllabic words and 64th for both. For all words this corresponds to the top 4.4% of the original vocabulary of 1478 words. Errors (i.e. inappropriate ranking for the correct word candidate) are due to alternate pronunciations not yet generated by the knowledge compiler (e.g., british pronunciation of CLASSES), sonorant classifier inaccuracies, inaccurate VUV-decisions, endpoint detection errors. Some of these problems can be improved by simply adding more speech knowledge to the compiler. We also hope that the addition of more filters will constrain acceptable word candidates further and reduce the effective subvocabulary.

6. Summary

In summary, I have demonstrated that prosodic cues can provide speech recognition systems with a powerful alternative perspective on the speech signal. A set of suprasegmental vocabulary filters was shown to constrain the possible word candidates in such a way, that the correct candidate was ranked on average 64th in a 1500-word large vocabulary. The filters operate knowledge intensive rather than search intensive, in order to allow for fast candidate preselection. Finally, all information used to describe a vocabulary entry is generated automatically, no human training is necessary when new vocabularies or new words are used.

1. T. Kaneko and N.R. Dixon, "A Hierarchical Decision Approach to Large-Vocabulary Discrete Utterance Recognition," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. 31, No. 5, October 1983, pp. 1061-1066.
2. D.W. Shipman and V.W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *ICASSP '82*, IEEE ASSP, 1982, pp. 546-549.
3. A.E. Rosenberg, L.R. Rabiner, J.G. Wilpon, D. Kahn, "Demisyllable-Based Isolated Word Recognition Systems," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. 31, No. 3, June 1983, pp. 713-726.
4. H.D. Hoehne, C. Coker, S.E. Levinson, L.R. Rabiner, "On Temporal Alignment of Sentences of Natural and Synthetic Speech," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. 31, No. 4, August 1983, pp. 807-813.
5. K. Prazdny, "Waveform Segmentation and Description Using Edge Preserving Smoothing," *Computer Vision, Graphics, and Image Processing*, Vol. 23, 1983, pp. 327-333.
6. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
7. K.S. Fu, *Syntactic Pattern Recognition and Applications*, Prentice-Hall, Inc., 1982.
8. G.D. Allen, "The Location of Rhythmic Stress Beats in English: An Experimental Study I and II," *Language and Speech*, Vol. 15 and 16, 1972, pp. 72-100 and 179-195.
9. N.J. Nilsson, *Learning Machines*, McGraw-Hill Book Company, 1965.
10. S. Makino, T. Kawabata, K. Kido, "Recognition of Consonant Based on the Perceptron Model," *ICASSP 83 Proceedings*, IEEE, 1983, pp. 738-741.
11. R.A Cole and R.A. Brennan, "A Pretty Good Formant Tracker," *The Journal of the Acoustical Society of America*, Vol. 74, 1983, pp. S15, (abstract only)
12. J. Allen, R. Carlson, B. Granstrom, S. Hunnicutt, D. Klatt, D. Pisoni, *Conversion of Unrestricted English Text to Speech*, Massachusetts Institute of Technology, 1979.
13. A.Waibel, "Towards Very Large Vocabulary Word Recognition," Tech. report 144, Carnegie-Mellon University Computer Science Department, 1982.