A Coarse Phonetic Knowledge Source for Template Independent Large Vocabulary Word Recognition

Helmut Lagger

Central Technology Division
Siemens AG, Munich, Germany

Alex Waibel

Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

In this paper we present a template independent knowledge source (KS), that uses coarse phonetic information to substantially constrain the candidate vocabulary for use in word hypothesization with very large vocabularies. It consists of three parts: the segmenter that breaks a test utterance up into a sequence of coarse phonetic classes, the knowledge compiler that generates a reference dictionary containing the appropriate coarse phonetic representations for each word candidate and finally, a matching engine. Coarse phonetic classification is performed using linear discriminant analysis, more specifically perceptron classification. The knowledge compiler first generates a phonemic representation and segmental durations by rule from a list of word candidates (i.e., from text), and then derives coarse phonetic class segments. Matching is performed by a nonlinear time alignment algorithm based on dissimilarity scores between detected and lexical coarse class segments. The coarse phonetic KS was tested by compiling a word list of approximately 1500 words. Using only the coarse classes Silence, Plosive, Fricative, Vocalic, Front Vowel, Back Vowel, Nasal and R, a vocabulary reduction to 5% of the original vocabulary is achieved at lower than 5% error rate for three different speakers.

1. Introduction

Most current speech recognition systems today cannot easily be extended to large vocabularies of several thousand words. Some of the most serious critical requirements that must be met by large vocabulary recognition systems are computational efficiency, practicality, flexibility and robust recognition accuracy. Searching a large vocabulary for word candidates must be done efficiently. Maintaining and collecting a database of reference word-templates becomes costly for large vocabularies and cannot be expected from the user of such a system. In addition, it is desirable to flexibly add or subtract new lexical items (dictionary entries) as the language or the needs of the user change over time. Finally, highly robust recognition algorithms must be developed to deal with the increasing acoustic similarity of words in a large vocabulary.

Several studies have proposed various methods to overcome some of the difficulties cited above. Recognition of smaller atomic units than the word, e.g., demisyllables [1] or phonemes [2] reduces or eliminates the required amount of user training. It has also been demonstrated that preselection of smaller subvocabularies could be achieved by means of relatively simple acoustic measures [3] given a reference dictionary of word templates. Shipman and Zue have shown that a large vocabulary can be reduced into surprisingly small subvocabularies if an error-free description of the utterance in terms of coarse phonetic classes is given [4, 5]. Alternate acoustic evidence such as suprasegmental cues in the signal were also shown to provide powerful constraints for search space reduction [6].

In the present work we extent these previous results and present and evaluate a knowledge source (KS) that achieves vocabulary reduction based on the detection of *coarse phonetic categories* and does not require excessive user training. It can be used to either preselect a smaller subvocabulary or to "raise activation levels" of various word candidates. *Template independence* and *flexibility* of the KS is achieved by a rule-based knowledge compiler that compiles an orthographic representation (text) of the candidate words into a coarse phonetic representation. The KS then compares the sequence of coarse phonetic categories in the incoming unknown utterance against the coarse class representation of each lexical item for satisfaction of its constraints.

In the following sections we first describe the classifier used to recognize sequences of coarse phonetic categories in a test utterance. We then discuss the knowledge compiler used to generate the reference dictionary and the matching engine that performs the recognition. Finally we present results of recognition experiments using a 1500 word

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.

CH2118-8/85/0000-0862 $1.00 © 1985 IEEE

vocabulary.

2. Coarse Phonetic Classifier

Proper classification of coarse phonetic categories is a critical step for the present knowledge source. Out of the many methods that could accomplish this task we have chosen to use linear classifiers, specifically, perceptrons, for their simplicity in automatic learning as well as recognition.

An incoming speech utterance is first sampled at a 16 kHz sampling rate and lowpass filtered by a 6.4-kHz antialiasing filter. We then compute a 256-point DFT every 3 ms over 20 ms frames of Hamming-windowed speech. The features used for classification consist of 54 spectral coefficients linearly spanning the spectral range as well as 8 coefficients spanning the spectral range logarithmically. In order to obtain a smoother contour of frame-by-frame coarse class decisions and alternatively to obtain a classifier that captures the typical time varying dynamic behavior of certain coarse classes (such as Plosives), three frames characterizing 15 ms of speech are combined into one feature vector. The resulting total of 186 features in such a feature vector contains redundant inter and intra frame information. In order to eliminate such redundant information principal component analysis is performed. An incoming feature vector is thus rotated by the eigenvectors of the covariance matrix derived from a set of more than 20,000 training feature vectors. Only the rotated features corresponding to the 54 largest eigenvalues are considered for further analysis.

Using the feature vectors obtained in this fashion, coarse phonetic categories can be detected using linear discriminant analysis, more specifically perceptron classifiers [7, 8, 9]. To train these classifiers, the rapidly converging relaxation method [9] was used. Fifty random utterances from a database of 1500 words, spoken in isolation by three different speakers were set aside as training data. Each of these 50 utterances was hand labelled according to the coarse phonetic classes, Silence, Fricative, Plosive and Vocalic, as well as additional labels for the vocalic parts, Nasal, Front, Back, R. Thus M, N, NG were labelled Nasal, front vowels (e.g., IY, IH, EH) were labelled Front, back vowels (such as AA, AO, UW) as well as the semivowel W and the glide L were labelled Back. The glide R, was labelled as its own category, R. This taxonomy was chosen empirically, so that good recognition performance could be achieved by the classifiers.

In a preliminary sorting step, all feature vectors corresponding to frames with the same coarse class label are collected in appropriate files. An error-correcting learning procedure (the relaxation method) produces for each coarse class a weight vector that defines a linear decision hyperplane used for classification. We obtain a total of 8 hyperplanes, each of which separates one of the classes mentioned above from all the others. Using the projections of all labelled data onto the axes normal to the decision hyperplanes, we compute for each class the probability of membership in the class as a function of distance along the normal using k_n-nearest neighbor estimation techniques [9]. Thus, the probability of membership in a given class is obtained by computing the scalar product of the weight vector and an incoming feature vector and looking up the corresponding probability in a table.

In the next step, segmentation into coarse phonetic segments and some post-processing is performed. All adjacent frames of speech data for which one of the perceptrons fires with clearly maximal probability are collapsed to one segment and assigned to one and only coarse phonetic class. Unclear regions are left undefined until subsequent postprocessing is performed. Context sensitive rules then attempt to determine the most likely identity of ambiguous segments. To eliminate unlikely segment sequences or to correct possible misclassifcations, higher level rules are applied. For example short final nasalized segments are eliminated, as well as short fricative segments caused by aspiration at the end of the utterance. Of course, such rules are applied conservatively to minimize the possibility of introducing extra errors. Further postprocessing breaks up long segments into smaller subsegments, yielding an average segment duration of approximately 30 ms. This is done to achieve optimal performance in the matching stage described below. Finally, the resulting segment class is encoded into one byte, specifying the identity of the segment. In order to gain greater computational efficiency we make "hard" segmental decisions and could potentially loose important information pertaining to lesser ranked candidates.

3. The Knowledge Compiler

The knowledge compiler takes a list of orthographic word candidates (written text) and automatically generates (a) a coarse phonetic representation and (b) segmental durations for each word. In the first step, parts of the MIT text-to-speech synthesis system [10] are used to hypothesize a phonemic representation and segmental durations for the word. The derived phonemic representation as well as the segmental durations are further processed in the second stage. First, consecutive segments belonging to the same coarse phonetic classes are collapsed. Conversely, diphthongs describing a move from back to front vowel or vice versa are split into corresponding coarse classes. Splitting into different coarse class segments is also done for phonemes such as EN, EM, ER, etc. Next, alternate pronunciations are derived by rule. This

is useful, for example, for reduced vowels (that are strongly influenced by context) and for phonemes that are close to coarse class category boundaries (for example, reduced schwa is near the decision boundary between front and back and tends to be heavily influenced by context). Alternate pronunciations are also useful for weak, voiced fricatives, such as the phoneme V. Finally, depending on the durations of the split or collapsed segments, long segments are broken up into smaller subsegments, resulting in an average segment duration of approximately 30 ms. This is done to ensure that segments in the unknown as well as the reference pattern will not differ too much in duration. This is important to achieve proper matching behavior.

4. The Matcher

The purpose of the matcher in this KS is to evaluate the degree of constraint satisfaction of individual lexical items with respect to the incoming sequence of coarse phonetic events. Since this involves searching a large corpus of lexical items, this evaluation must be performed efficiently. Using coarse phonetic classes greatly reduces computational cost by requiring matching to be done on short sequences of segment labels only. To allow for missed or extra segments, matching is performed using a nonlinear time alignment algorithm. The dynamic programming algorithm proposed by Itakura [11] was chosen for this task. A critical design consideration is the choice of the distance metric. It must be simple to compute and provide the discriminatory information we are seeking. Since in our case, only 8 distinct coarse phonetic classes are possible, all possible class-to-class distances can be easily precompiled into a look-up table for efficient evaluation. These class-to-class distances were derived empirically from training data incorporating some general heuristics. For example, distances between Fricatives, Silences, and Vocalic segments receive greater weight than distances between the vocalic classes Front, Back, Nasal and R. Thus, substantial computational savings can be achieved as well as greater flexibility in defining the distances themselves. In addition, alternate pronunciations can be taken care of simply by appropriate definitions in the distance table. Variations in temporal behavior are, of course, corrected by the warping algorithm itself. The algorithm recovers gracefully from missed or extra segments by means of the dynamic programming alignment, while segment confusions become non-fatal through the use of alternate pronunciations and appropriate class-to-class distance values. Note that an inaccurate sequence of coarse phonetic classes will therefore still lead to acceptable lexical retrieval, despite the fact that hard decisions were made at the segmental level.

5. Performance Evaluation

For the recognition experiments reported below, a vocabulary of 1478 words was compiled from two word lists containing the 900 most frequent written and 900 most frequent spoken words in English [12, 6]. Three male American speakers (MSH, MRN and MKD) uttered the entire word list once. For each speaker, 50 randomly selected utterances were set aside to develop the rules for the knowledge compiler and to train the classifiers as described above. The knowledge compiler was run over the entire word list of 1478 words and thus includes 1428 "new" words. The coarse phonetic KS was then tested for each speaker on 500 words randomly selected from this set of 1428 "new" words. The results of this experiment are given in Figure 1. The three curves show for each speaker the recognition score in percent as a function of the number of top candidates included, up to 250. The graph shows that using only the coarse classes Silence, Fricative, Plosive, Vocalic, Front Vowel, Back Vowel, Nasal and R Glides the right word candidate is included in the top 5% of the vocabulary (~ 75 words) more than 95% of the time and in the top 17% of the vocabulary (~ 250 words) more than 99% of the time. 37.6% of all the utterances were identified uniquely as first choice candidates out of the 1500 word vocabulary.

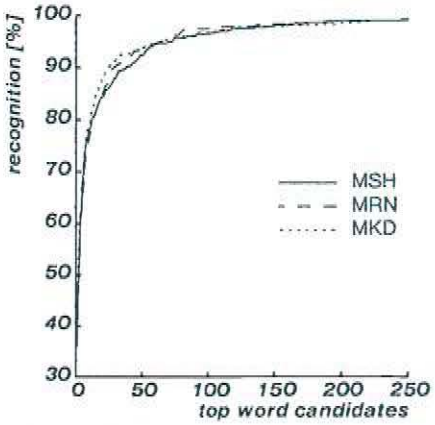

Figure 1. Recognition Performance for a 1500 Word Vocabulary

What are the theoretical limits on performance one might expect from the present approach? We have pooled all lexical items that match perfectly with each other into groups and obtained group sizes ranging between 1 and 17. The expected group size [12] was 2.8. This result is comparable to statistics reported by Shipman and Zue [4]. This would suggest that given errorfree coarse phonetic input an average rank of 1.4 should be expected for the correct word hypothesis. The present recognition results indicate an average rank of 6.9. The differences

between this theoretical upper bound and the recognition results are due in part to compiler errors, i.e., inaccurate coarse phonetic descriptions provided by the compiler, or alternate pronunciations that were not anticipated by the compiler. A second source of error leading to reduced discriminability is given by classifier errors and/or the variabilities found in human speech including spurious speaker-generated noise (such as pops, clicks, lipsmacks) frequently resulting in endpoint detection errors, aspiration noise at the end of utterances, nasalization of vowels and the like. Further improvements towards classification of acoustic events and further careful study of the possible acoustic manifestations of English words for better compiler rules might improve these results.

6. Summary

In summary, we have presented a knowledge source for template independent large vocabulary word recognition. The KS uses only coarse phonetic classes and does not require extensive user training. All lexical information needed for recognition is automatically generated from text. For a 1500 word vocabulary it will include the correct word candidate among the 75 (~5% of vocabulary) best candidates with an error rate of less than 5%. Such a KS is useful to either preselect a smaller subvocabulary or as an independent KS aimed at "raising activation levels" for individual word candidates. We believe that in a distributed cooperative arrangement together with prosodic, lexical, fine phonetic and coarse phonetic KSs, a small set of word hypotheses can be obtained efficiently for Large Vocabulary Speech Understanding Systems.

References

1. 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.
2. S. Makino and K. Kido, "A Speaker Independant Word Recognition System Based on Phoneme Recognition for a Large Size (212 Words) Vocabulary," *ICASSP '84 Proceedings*, IEEE, 1984, pp. 17.8.1-17.8.4.
3. 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.
4. D.W. Shipman and V.W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *ICASSP '82 Proceedings*, IEEE, 1982, pp. 546-549.
5. D.P. Huttenlocher and V.W. Zue, "A Model of Lexical Access Based on Partial Phonetic Information," *ICASSP '84 Proceedings*, IEEE, 1982, pp. 26.4.1-26.4.4.
6. A. Waibel, "Suprasegmentals in Very Large Vocabulary Isolated Word Recognition," *ICASSP '84 Proceedings*, IEEE, 1984, pp. 26.3.1-26.3.4.
7. S. Makino, T. Kawabata, K. Kido, "Recognition of Consonant Based on the Perceptron Model," *ICASSP 83 Proceedings*, IEEE, 1983, pp. 738-741.
8. N.J. Nilsson, *Learning Machines*, McGraw-Hill Book Company, New York, NY, 1965.
9. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, 1973.
10. J. Allen, R. Carlson, B. Grandstrom, S. Hunnicutt, D. Klatt, D. Pisoni, *Conversion of Unrestricted English Text to Speech*, Massachusetts Institute of Technology, Cambridge, MA, 1979.
11. F.Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-23, No. 1, February 1975, pp. 67-72.
12. A.Waibel, "Towards Very Large Vocabulary Word Recognition," Tech. report 144, Carnegie-Mellon University Computer Science Department, 1982.