A Procedure to Generate Training Sequences for a Connected Word Recognizer using the Segmental *k*-means Training Algorithm

R. P. Mikkilineni
J. G. Wilpon
L. R. Rabiner

AT&T Bell Laboratories
Speech Research Department
Murray Hill, New Jersey 07974

ABSTRACT

Past research has shown that a connected digit recognition system, based on either word templates or word hidden Markov models (HMM), could effectively be trained using a segmental *k*-means training procedure. In these studies, a set of randomly generated digit strings of variable length was used to train the recognizer. However, problems were encountered when this training procedure was extended to systems with medium to large vocabularies.

For a training set to be effective, it should represent each vocabulary word and its acoustic variability within the context of all valid input strings defined by a task dependent grammar. In this paper, we propose a procedure to generate a training set of sentences with these properties. Using this procedure, a training set of sentences was generated for a connected word recognition system simulating an airline flight reservation task.

Several speaker dependent automatic speech recognition (ASR) experiments were performed to assess the effectiveness of the training set generated using the new procedure. The results of these experiments showed that the string accuracy was about 98% when tested on independent sets of test sentences for the three talkers.

1. Introduction

A number of approaches to the problem of training a connected word recognizer using word templates or hidden Markov models (HMM), have been proposed and evaluated in earlier experiments [1-4]. The earliest training procedures used isolated words to produce reference patterns [1-2]. This worked reasonably well for users who carefully articulated their test strings. However, the performance of the recognition system degraded to unacceptable levels for naturally spoken continuous speech. This was because the isolated words did not represent the co-articulation between adjacent words of a connected word string. Additionally, the duration of a word spoken in isolation was often vastly different from its duration in a continuously spoken word string.

To overcome this problem, an embedded word training procedure was developed for digit sequences and for spoken spelled letters. In this procedure, word tokens were extracted from continuously spoken three-digit strings. These tokens were then combined with isolated digit patterns to produce a set of reference patterns for each vocabulary word. This procedure was successfully used to improve reference patterns for connected digit recognition and for spelled letter recognition of names from a directory.

A continuous speech training procedure, called the segmental *k*-means training algorithm was developed next [3]. This procedure created word reference patterns from continuously spoken speech via an iterative segmentation and modeling procedure. The training sentences were chosen at random from the set of all syntactically valid sentences for the recognition task. For example, several hundred random digit strings were used to train a connected digit recognizer. The string accuracy of connected digit strings of known length was shown to be about 99% in a speaker trained mode even when the articulation rate was over 170 words per minute.

The segmental *k*-means training procedure was evaluated on the task of simulating an airline flight reservation. This system had a vocabulary of 127 words and the set of syntactically valid sentences were represented by a finite state network. The training set consisted of five repetitions of each of 52 sentences, which were selected manually such that all the vocabulary words and the states of the finite state grammar were represented. This resulted in a string accuracy of 93% (speaker trained mode) when tested on an independent test set of sentences. On the other hand, the recognizer had an accuracy of 99% when tested on the training set of sentences. From these results, it is clear that the training set of sentences did not adequately characterize the words in the vocabulary for all possible contexts in which they could be spoken.

In this paper, we propose a procedure to generate a more representative set of sentences for training a connected word recognition system. Using this procedure, a set of sentences was generated to train the airline flight reservation system. The effectiveness of this training set was then evaluated through a series of experiments in the laboratory.

In section 2, we present the procedure used to generate the set of training sentences based on syntax specification via a finite state grammar. In section 3, we describe the application of the training procedure to the airline reservation system. Finally, in sections 4 and 5, we present the results from a series of connected word recognition experiments using both a template-based and an HMM-based ASR system.

2. Selection of the Training Set

One of the important issues, in the design of a connected word recognition system, is the selection of the text (i.e. the set of sentences) for training the recognizer. The set of all valid input strings to the system is undoubtedly the best training set. However, this set is often excessively large. For example, for connected digit recognition with no syntactic constraints and a valid string length of up to 6 digits, the complete training set has over one million valid input strings. Consequently, it is essentially never practical to train the recognition system for all valid inputs. Therefore, one has to choose a fraction of all valid sentences, which adequately represents the language as specified by the finite state grammar, to train the recognizer. The effectiveness of the training set is measured by the accuracy of the recognition system on an independent test set of sentences.

2.1 Statement of the Problem

Let V, G, and L represent the vocabulary, grammar and the language of the recognition task. The problem of selecting a training set can be stated as follows. Given V, and G, select a subset, T of L such that a connected word recognizer, trained using T, exhibits

CH2561-9/88/0000-0433 $1.00 © 1988 IEEE

comparable string accuracies when tested on T and on any other representative subset of the language, L.

In this paper, G is defined to be a finite state grammar which is represented as a directed graph with no cycles. An example of such a graph is shown in Figure 1. For this type of grammar, the number of valid sentences is finite and can easily be enumerated.

2.2 Procedure to Generate a Training Set of Sentences

In a recognition system, the acoustic variability of each vocabulary word is represented by the word reference patterns. As these patterns are derived from a training set, it is reasonable to assume that a good training set should represent the acoustic variability of each vocabulary word in the context of a given language. Thus, the training set of sentences should represent each vocabulary word in all possible contexts of the grammar.

To generate such a training set, a sequence of training sets with increasing representation of the words in various contexts of the grammar is defined and an algorithm is presented to generate such a training set of sentences. Let $T_0, T_1, T_2, \ldots, T_n$, where n is the maximum length of a sentence in the language generated by the grammar, be a sequence of training sets. The set, T_0 represents a training set with isolated words only. The sets, T_1 to T_n, are sets of sentences with the following properties. T_1 represents all the edges (and, consequently, the words associated with them) of the directed graph associated with the grammar. The set, T_2 has all the properties of T_1 as well as the property that it represents all the valid edge pairs in the graph. The set, T_3 has all the properties of T_2, and all the valid edge triplets in the graph. Continuing this process, we find that the training set, T_n contains all the sentences in the language.

The last training set, T_n represents each vocabulary word in all possible contexts. Thus, it represents all the acoustic variability associated with all the words. But, since the number of valid sentences is exceedingly large for any reasonable task, it is not practicable to use T_n as a training set. At the other extreme, T_0 does not represent any acoustic variability associated with co-articulation between words and it has been shown to be ineffective for training a connected word recognition system. The first reasonable training set which contains some representation of the acoustic variability associated with the vocabulary words, is T_1.

We present a procedure to generate the training set T_1 given a directed graph representation of the task grammar. Let M represent the minimum frequency, of each vocabulary word, required in the training set. Then the algorithm for generating T_1 is as follows:

```
set $T_1$ = null
set $f_w$ = 0 for each vocabulary word
set $f_a$ = 0 for each arc of the graph
for each arc in the graph
do
  if ($f_a$ = 0 for the current arc, or, $f_w < M$
          for the word associated with the current arc)
  do
    trace back a path segment from the current arc to start node
    trace forward a path segment from the current arc to
       terminal node
    combine these path segments to produce a complete path
       and generate the sentence associated with that path
    increment $f_w$ for all words in the sentence
    increment $f_a$ for all arcs in the path
    add this sentence to $T_1$
  end do
end do
```

When this algorithm terminates, the set T_1 contains the desired set of training sentences. This algorithm does not generate the minimal set of sentences covering all the arcs. In tracing a path segment, the arc with the smallest f_a is selected if more than one arc is available at any stage. It should be noted that this basic algorithm may be readily extended to generate any training set, T_i.

3. Application to Airline Reservation System

3.1 Vocabulary and Grammar

The task to which the new training procedure was applied was an airline flight reservation system [5]. The vocabulary contained 127 words typically used to make airline database inquiries. The valid sentences are represented by a finite state grammar with 144 states, and 450 transitions. Each transition represents a vocabulary word and a sequence of transitions from the initial state to a terminal state defines a valid sentence. The maximum entropy of the language is 2.15 bits/word. There are over 6 billion valid sentences in the language generated by this grammar. They range in length from 4 to 22 words.

3.2 Experimental Setup

A block diagram of the syntax directed connected word recognition system used in this study is shown in Figure 2. A detailed description of each module in this figure has been reported previously [3-5]. As such, we will present only a brief review of the function of each module.

3.2.1 Feature Extraction The input speech signal is recorded off a local, dialed-up, telephone line and sampled at a 6.67 kHz rate. The digitized signal is then pre-emphasized by a first order digital network (with a pre-emphasis factor of 0.95) and then blocked into frames of 45 msec. Consecutive frames are spaced 15 msec apart. A Hamming window is used on each frame and an 8-th order linear predictive coding (LPC) analysis is performed. In the recognition process, the LPC feature set is transformed to cepstral and differential cepstral features [6].

3.2.2 Reference Pattern Module The reference pattern module contains the reference patterns for each vocabulary word. They are either word templates or hidden Markov models for each vocabulary word. The following sets of reference patterns were used in the experiments.

a. speaker independent (SI) reference patterns based on isolated words,

b. speaker dependent (SD) reference patterns based on isolated words,

c. SD reference patterns based on connected words from the new training procedure.

3.2.3 Syntax Module The syntax module describes the syntactic constraints on the vocabulary words. These constraints are specified by a finite state grammar which is represented as a directed graph. Each node of the graph represents a state and each arc is associated with a vocabulary word [5]. For the airline reservation task, this graph contains 144 nodes, and 450 transitions.

The main purpose of this module is to limit the search space of the recognizer to the language generated by the grammar.

3.2.4 Recognition Algorithm The recognition algorithm is a pattern matching or model matching algorithm depending on whether templates or hidden Markov models (HMM) are used as reference patterns. It uses the level building algorithm [7] to determine the best sequence of reference patterns that matches a given connected word string. The search procedure limits the search to valid sentences of the language as generated by the syntax. The output of the system is the best syntactically valid sentence from among all possible sentences.

3.3 Training and Test Sets

Using the algorithm described in section 2.2, we generated the training set T_1. This set consisted of 265 sentences. Each subject recorded the training sets T_1 and T_0 (which consisted of 5 isolated utterances of each vocabulary word). The training set, T_0 was used to estimate the string accuracy of the connected word recognition system trained with isolated words only. Further, the reference patterns produced from T_0 were used to initialize the segmental k-means training procedure to train the system with T_1.

In addition to these SD training sets, a speaker independent set of reference patterns was used in this study. They were generated from isolated word utterances of each vocabulary word by 100 talkers [8].

The test set consisted of 250 sentences and contained at least one occurrence of each vocabulary word. Further, it represented all the nodes of the finite state grammar at least 5 times.

The anticulation rates for the training and test sets range from 111.1 to 337.4 words per minute.

4. Recognition Results Using a Template-Based ASR System

4.1 Isolated Word Training

The results of the experiment to evaluate the performance of the system using isolated word training are presented in this section. Two sets of templates, one speaker dependent (SD) and the other speaker independent (SI) were used in this experiment. For the SD templates, the five tokens of each vocabulary word from the training set, T_0, were clustered using the modified k-means clustering algorithm [8] to produce three templates per word. The SI templates, consisting of 6 templates per word, were produced by clustering 100 tokens of each word spoken by 100 different talkers.

The average sentence accuracy was 71.8% using the SI templates and 87.1% using the SD templates. The word accuracies using the SI and SD template sets were 93.6% and 97.6% respectively.

4.2 Connected Word Training Using SD Reference Patterns

This template set, for the SD case, was generated using the segmental k-means training algorithm applied to the training set T_1. The k-means loop was initialized using the isolated word templates (for each talker) defined in section 4.1. Using this procedure, three templates per word were generated from the segmented continuous speech using clustering techniques. The results from these recognition experiments are presented in Table 1. The results include the string accuracies from the recognition tests on the training set and an independent test set. The average sentence accuracy was 99.2% on the training set. This indicates that the templates represented the training data adequately. The system had a sentence accuracy of 98.1% when tested on the independent test set. For speaker 1, there were 8 word substitution errors. There were 1 substitution and 2 word insertion errors for speaker 2. For speaker 3, there were 2 substitution and 1 insertion errors.

This low error rate on the training data as well as on the independent test set indicates that the procedure to generate training set material succeeded well in representing the acoustic structure of the vocabulary from the task grammar using all edges of the grammar network.

4.3 Connected Word Training Using SI Reference Patterns

In this experiment, our goal was to determine if we could initialize the segmental k-means training procedure from a set of speaker independent templates rather than from the SD isolated word set, T_0. In this case, the user was not required to record the set of isolated words, T_0, in addition to the training set, T_1. The results of the experiments using SI initialization are given in Table 1. The average sentence accuracies on the training and test sets were 99.3% and 98.3% respectively. These results are not significantly different from those results generated using SD isolated word templates for initialization of the segmental k-means loop. These results indicate that an effective set of SD templates could be produced using SI template initialization of the segmental k-means training procedure.

5. Recognition Results Using HMM-Based ASR System

The hidden Markov models used to characterize individual words are first order, left-to-right, Markov models with 5 to 10 states. Transitions between words are handled by a switch mode from the last state of one word model, to the first state of another word model, in the level building implementation.

5.1 HMM Recognition From Isolated Word Training

The results on sentence accuracy for the HMM recognition tests, based on isolated word training, are given in this section. The size of the models was 8 states, 9 mixtures per state for talkers 1 and 3 and 5 states, 3 mixtures per state for talker 2. For models derived from the SI isolated words, the average sentence accuracy of the HMM recognizer is essentially the same as that of the template-based recognizer, namely about 72%. For T_0, the avearge sentence accuracy is 85.9%.

5.2 HMM Recognition from SD Bootstarpped Reference Patterns

The results on sentence accuracy for the HMM recognizer based on T_1 training, bootstrapped from T_0 (i.e. SD isolted words) are given in Table 2. For the training set there were no errors for any of the talkers. For the independent test set, the average string accuracy was high (98.5%) and again was comparable to that obtained in the template-based approach. For these results models with; 12 states, 5 mixtures per state were used for talker 1; 8 states, 5mixtures per state for talker 2; and 10 states, 5 mixtures per state for talker 3.

5.3 HMM Recognition from SI Bootstrapped Reference Patterns

The results on sentence accuracy for the HMM recognizer based on T_1 training, bootstrapped from SI isolated word reference models are given in Table 2. Again there were no errors on the training set of sentences, and the performance on the independent test set was essentially the same as from a template-based approach, or from the HMM-based system with models bootstrapped from SD isolated word training.

6. Summary

We have described an effective way of training a connected word recognizer based on using a representative set of continuous speech training sentences generated from a finite state network which characterizes the syntax of the recognition task. It was shown that when this training procedure was applied to the task of recognizing continuous sentences from an airline reservations task (127 word vocabulary, entropy of 2.1 bits /word), in a speaker trainind mode, average sentence accuracies on the order of 98-99% were obtained using either word templates or statistical models.

References

1. H. Sakoe, "Two Level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 6, December 1979, pp. 588-595.
2. J. S. Briddle, M. D. Brown, and R. M. Chamberlain, "An Algorithm for Connected Word Recognition," Automatic Speech Analysis and Recognition, edited by J. P. Haton, D. Riddle Publishing Co., Dordrecht, Holland, 1982, pp. 191-204.
3. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Segmental k-means Training Procedure for Connected Word Recognition," AT&T Technical Journal, Vol. 65, Issue 3, May 1986.
4. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Model-Based Connected-Digit Recognition System Using either Hidden Markov Models or Templates," Computer Speech & Language, Vol. 1, number 2, December, 1986.
5. S. E. Levinson, and L. R. Rabiner, "A task-Oriented Conversational Mode Speech Understanding System," Speech and Speaker Recognition, edited by M. R. Schroeder, S. Karger AG., Basil, Switzerland, 1985, pp. 149-196.
6. L. R. Rabiner, J. G. Wilpon, F. K. Soong, and A. E. Rosenberg, " High Performance Connected digit Recognition using hidden Markov Models," submitted for publication.

7. C. S. Myers and L. R. Rabiner, "Connected Digit Recognition Using a Level Building DTW Algorithm," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-2, June 1981, pp. 351-363.

8. J. G. Wilpon, and L. R. Rabiner, "A Modified k-Means Clustering Algorithm for Use in Isolated Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, June 1985, pp. 587-594.

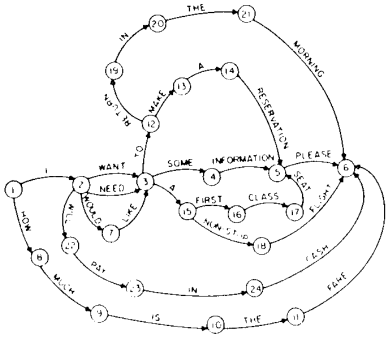

Figure 1 : An Example of a Finite State Grammar

Speaker	SD Word References		SI Word References	
	Training Set	Test Set	Training Set	Test Set
1	98.1	96.8	98.8	97.6
2	99.6	98.8	100.0	98.8
3	100.0	98.8	99.2	98.4
Average	99.2	98.1	99.3	98.3

Table 1. Recognition Accuracy with Templates Derived from Connected Word Training Using SD and SI Isolated Word References.

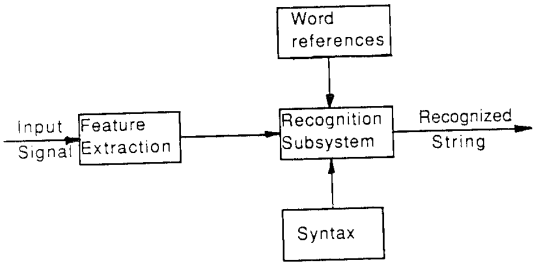

Figure 2 : An Overview of a Syntax Directed Connected Word Recognition System

Speaker	SD Word References		SI Word References	
	Training Set	Test Set	Training Set	Test Set
1	100.0	98.8	100.0	99.6
2	100.0	97.6	100.0	97.2
3	100.0	99.2	100.0	98.0
	100.0	98.5	100.0	98.2

Table 2. Recognition Accuracy with HMMs Derived from Connected Word Training Using SD and SI Isolated Word References.