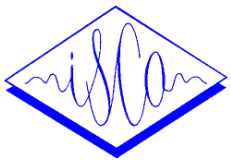

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# MINIMUM ERROR RATE TRAINING OF INTER-WORD CONTEXT DEPENDENT ACOUSTIC MODEL UNITS IN SPEECH RECOGNITION

*W. Chou, C.-H. Lee and B.-H. Juang*

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill,NJ 07974, U.S.A.

## ABSTRACT

In this paper, we study the issues related to string level acoustic modeling in continuous speech recognition. A new approach based on the minimum string error rate criterion is proposed to the training of inter-word context dependent acoustic model units. Under the proposed approach, the inter-word context dependent acoustic model units are modeled at the global string level by directly applying the minimum string error rate based discriminative analysis to string level acoustic model matching. Experimental results indicate that a significant error rate reduction can be achieved through the proposed approach. Based on the proposed approach, the best performance obtained by a gender-independent model on the TI connected digit corpus is 0.24% word error rate and 0.72% string error rate.

## 1. INTRODUCTION

One of the key issues in continuous speech recognition is how to select and model the basic speech recognition units for recognition. These basic speech recognition units can be based on the phonemes or based on the actual words in the vocabulary. In order to cope with the acoustic variations of the speech signal, various context dependencies are introduced into the structure of these basic speech recognition units. In particular, inter-word context dependent model units are widely used in speech recognition.

The introduction of inter-word context dependency in speech recognition has a significant impact on the acoustic modeling accuracy, and leads to a remarkable improvement in speech recognition performance. With the use of inter-word context dependent model units in speech recognition, the acoustic events at the word junctions, such as coarticulations, can be modeled more precisely than using only the context independent units or the units whose context dependencies are specified within the boundaries of words[9].

However, more accurate training procedures are needed in order to estimate the parameters for these high resolution acoustic model units. Recently, the minimum error rate based pattern recognition approach has become increasingly popular in speech recognition [1]-[6]. In a string model based approach to minimum error rate training [4], the model parameters of the HMMs used in speech recognition are estimated at the string level according to the global string model matching. In this paper, we extend the string model based minimum string error rate training approach to the training of inter-word context dependent model units in speech recognition. We will show in this paper, the new string model based minimum error rate training procedure can be applied to inter-word context dependent model units. Comparing with the conventional maximum likelihood (ML) approach, this new training procedure leads to further performance improvement for using inter-word context dependent model units in speech recognition.

In a connected digit recognition experiment on TI connected digit corpus using only one gender-independent model with inter-word context dependent model units, we achieved a string error rate of 0.72% and a corresponding word error rate of 0.24% with unknown length decoding. In comparison with the model obtained through the conventional ML approach, the string and word error rate reduction achieved through minimum error rate training is over 25%.

The organization of this paper is as follows. We introduce the formulation of minimum string error rate training in Section 2. We discuss issues related to training inter-word context dependent model units in Section 3. Experimental results are presented in Section 4.

## 2. STRING LEVEL ACOUSTIC MODELING

One assumption made in continuous speech recognition is that the acoustic model of a word string is formed by a linear concatenation of the basic speech recognition unit models according to a given lexical transcription. However, one of the distinct features in continuous speech recognition is that word errors committed by the recognizer during the process of recognition are based on the string level acoustic model matching. The decoding process in continuous speech recognition is to compare (implicitly) all possible string models at the whole utterance level. The word string whose string model has the highest likelihood score is chosen as the decoded string. Therefore, training procedures for basic speech recognition model units should be aimed at improving the acoustic resolution at the global string level

so that the recognition error rate can be reduced.

In continuous speech recognition using hidden Markov models (HMMs), each basic speech recognition unit is modeled by an HMM with the model parameters to be estimated from the training data. Let $O = \{O_1 \cdots O_T\}$ be the feature vector sequence extracted from the speech utterance. The string model for the word sequence $S = w_1 \cdots w_n$ given $O$ is the one which describes the word string $S$ and best matches the input speech utterance. This is typically determined by a Viterbi alignment, because many possible string models can describe the same word string, depending on the silence option, the optional pronunciations, phonological rules and so forth. The string model for the word string $S$ is defined by

$$\bar{M}_S = \underset{M_S}{\text{argmax}} \log f(O, \Theta_{M_S} \mid \Lambda), \tag{1}$$

where $M_S$ is a possible string model for $S$, $\Lambda$ is the set of the basic speech recognition model units, $\Theta_{M_S}$ is the optimal state sequence from Viterbi alignment in the string model $M_S$, and $\log f(O, \Theta_{M_S} \mid \Lambda)$ is the log-likelihood score along the best state sequence $\Theta_{M_S}$.

In the approach proposed in [4], the top $N$ competing string models are incorporated in training to characterize the string level acoustic model matching in recognition. The top $N$ competing string models are obtained from $N$-best decoding. We first identify the top $N$ word string hypotheses, and then, given each word string hypothesis, the string model can be uniquely determined according to equation (1). The top $N$ best string hypotheses $\{S_1, \cdots, S_N\}$ are defined inductively as follows,

$$S_1 = \underset{S}{\text{argmax}} \log f(O, \Theta_{\bar{M}_S}, S \mid \Lambda), \tag{2}$$

$$S_k = \underset{S \neq S_1, \cdots, S_{k-1}}{\text{argmax}} \log f(O, \Theta_{\bar{M}_S}, S \mid \Lambda). \tag{3}$$

The top $N$ string models are embedded into a specially designed loss function. The problem of "optimal classifier" design and the problem of the parameter estimation for the models of the speech recognition units become of finding the right parameter set to minimize the "sample risk" defined as the average cost incurred in classifying the set of the training design samples. The loss function in this approach is constructed through the following four steps.

(1) Discriminant function in minimum string error rate training is defined as

$$g(O, S_k, \Lambda) = \log f(O, \Theta_{\bar{M}_{S_k}}, S_k \mid \Lambda), \tag{4}$$

where $S_k$ is the $k$-th best string, $\Theta_{\bar{M}_{S_k}}$ is the optimal path (state sequence) of the $k$-th string given the model set $\Lambda$, and $\log f(O, \Theta_{\bar{M}_{S_k}}, S_k \mid \Lambda)$ is the related log-likelihood score on the optimal path of the $k$-th string.

For the correct string $S_{lex}$, the discriminant function is given by

$$g(O, S_{lex}, \Lambda) = \log f(O, \Theta_{\bar{M}_{S_{lex}}}, S_{lex} \mid \Lambda), \tag{5}$$

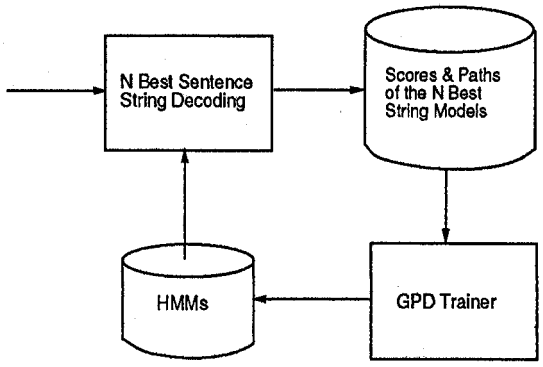


**Figure 1:** Diagram of string model based minimum string error rate training.

where $S_{lex}$ is the correct string, $\Theta_{\bar{M}_{S_{lex}}}$ is the optimal alignment path and $\log f(O, \Theta_{\bar{M}_{S_{lex}}}, S_{lex} \mid \Lambda)$ is the corresponding log-likelihood score.

(2) Misclassification measure in minimum string error rate training is defined as

$$d(O, \Lambda) = -g(O, S_{lex}, \Lambda) + \log\{\frac{1}{N-1} \sum_{S_k \neq S_{lex}} e^{g(O, S_k, \Lambda)\eta}\}^{\frac{1}{\eta}}. \tag{6}$$

(3) Loss function in minimum string error rate training is defined as

$$l(O, \Lambda) = \frac{1}{1 + e^{-\gamma d(O, \Lambda)}}, \tag{7}$$

where $\gamma$ is a positive constant, which controls the slope of the sigmoid function.

(4) The expected loss which is associated with the string error rate is given by

$$L(\Lambda) = E_O[l(O, \Lambda)]. \tag{8}$$

The model parameter estimation is to minimize the expected loss (8) which relates to the string error rate minimization. This can be achieved by using a sequential procedure according to the generalized probabilistic descent algorithm

$$\Lambda_{n+1} = \Lambda_n - \epsilon_n U_n \nabla l(O, \Lambda), \tag{9}$$

where $\epsilon_n$ is a sequence of step size parameters, $U_n$ is a sequence of positive definite matrices[6].

## 3. ISSUES RELATED TO INTER-WORD CONTEXT DEPENDENT MODEL UNITS

Speech recognition is a decision theory problem which is to determine the best sequence of words $S = w_1 \cdots w_n$ that maximizes

$$Pr(w_1 \cdots w_n) Pr(O_1 \cdots O_T \mid w_1 \cdots w_n), \tag{10}$$

where $Pr(w_1 \cdots w_n)$ is the probability of the word sequence $w_1 \cdots w_n$ from the language model, and $Pr(O_1 \cdots O_T \mid$

$w_1 \cdots w_n$) is the conditional probability from the acoustic model. Without inter-word context dependent model units, the distribution of each word in the sequence is assumed independent from each other, and the conditional probability from the acoustic model is approximated by the direct product of the conditional probability of each individual word model.

A string model incorporating inter-word context dependent acoustic model units is more accurate and the probability distribution of each word in the string becomes conditioned upon the joint acoustic events of its neighboring words. With inter-word context dependent model units, the basic speech recognition units in the string model are "glued" together through the context dependency which extends across word boundaries. This particular structure is actually more consistent with the criterion used in the string model based minimum error rate training than the modeling structure of context independent or intra-word context dependent model units, because the string level modeling constraint is explicitly defined in the definition of the inter-word context dependency.

With inter-word context dependent model units, each word model has multiple fan-in heads at the beginning of the word and multiple fan-out tails at the end of the word. This leads to different topology in the word model. For example, when using phone based speech recognition units, a three phone word is modeled by the head units, body units and tail units; a two phone word does not have a body and all possible head units of the word will be merged directly with the tail units. Moreover, a one-phone (or mono-phone) word is, in all its possible surrounding contexts, represented by a collection of inter-word context dependent model units whose central phone corresponds to the word itself. Therefore, the context dependency of a one-phone word will cross two word junctions involving three words in the word string.

This detailed specification of inter-word context dependent model units makes it difficult to accurately identify the top $N$ string models in training. The top $N$ string models needed for minimum error rate training become extremely fragile to the inaccuracies or errors incurred during the $N$-best decoding process. Inaccurate string level modeling will have an adverse effect on the acoustic resolution of the model units being built during the process of training. In order to solve this problem, we implement a high resolution $N$-best decoding scheme[7] in which the inter-word context dependencies are exactly preserved among the top $N$ competing string models obtained for training. This new search scheme is based on a forward partial path map preparation and a backward tree search. The backward tree search follows the context dependency and the language model used in the forward search and can trace back multiple word junctions, so that the context dependency can be handled exactly even for one-phone words.

One of the important issues in acoustic modeling is how to model the unseen word strings. The coverage in the training material on possible word strings is always very limited. This situation becomes even more acute for acoustic modeling using inter-word context dependent model units, because they are more dependent on the string level information and more sparse in the training material. In string model based minimum error rate training, this situation is improved by incorporating confusable word strings from $N$-best decoding in training. These confusable word strings are based on the same utterance in the training set, but have different word contents from the correct word string. Moreover, they are acoustic driven and not limited to be the word strings in the original training set. As a consequence, some of them can be strings "unseen" in the training set and thus, contain new string level word contexts.

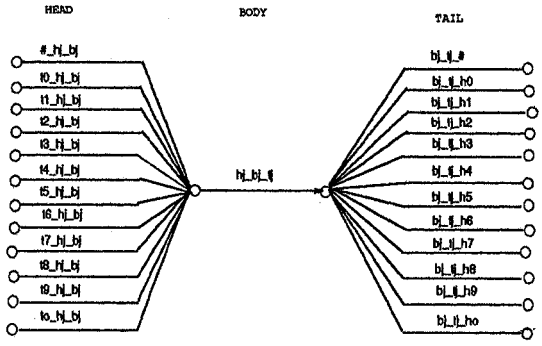


**Figure 2:** Diagram of digit model with acoustic inter-word context dependent model units.

Therefore, in addition to the new training criterion used in the string model based minimum error rate training, the training material used in model parameter estimation is also driven by the acoustic model matching at the string level. The competing string models used in training are based on the acoustic resolution of the basic speech recognition model units. Word strings which are confusable with the correct word string for the same training speech sample will be modeled discriminatively. This unique property of minimum string error rate training is more important for inter-word context dependent model units, because under this new training approach, these word junction model units can be estimated at the global utterance level based on an expanded set of word contexts from the competing string models.

## 4. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the proposed approach, we choose a connected digit recognition task and use a gender-independent model in which 95% of the units in the model are acoustic based inter-word context dependent model units. For such a high resolution model, the requirement for the acoustic modeling accuracy is extremely high[8]. The acoustic based inter-word context dependent model units are defined directly according to the acoustic events at the word junctions. Unlike other types of inter-word context dependent model units, it does not require a

| Training Method 8700 sents. | Word (Error_rate) | String (Error_rate) |
|---|---|---|
| Baseline (ML) | 0.33% (93) | 0.97% (84) |
| Str_Err_GPD | 0.24% (70) | 0.72% (63) |
| Err_Reduction | 26% (23) | 25% (21) |

Table 1: Performance comparison of an model obtained from the string model based minimum string error rate training.

detailed specification of the linguistically based phone lexicon and the associated phonological rules. In the case of connected digit recognition, each digit $j$ is divided into three parts, namely the head unit $h_j$, the body unit $b_j$ and the tail unit $t_j$. The head and tail units are inter-word context dependent units. Each head part of the digit are split into two types of representations, $\#h_jb_j$ and $t_ih_jb_j$, depending on the preceding context being a silence or a digit. Similarly, the tail part of each digit is split into two types of representations, $b_jt_j\#$ and $b_jt_jh_k$ depending on the following context being a silence or a digit $k$. Thus, the topology of a digit model has 12 fan-in heads and 12 fan-out tails as illustrated in Figure 2.

Our system in the experiments was based on the bandpass filtered (from 100Hz to 3.8KHz) speech signal obtained from the original 20KHz wideband speech in TI connected digit corpus. The bandpass filtered signal was down sampled at 8KHz and passed through a pre-emphasis filter to flatten the signal spectrally. A 30 msec Hamming window with 10 msec shift was used. A 10-th order auto-correlation analysis was performed to compute the LPC derived cepstral analysis. The feature vector used in recognition has 39 parameters, including 12 liftered cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, the normalized energy, and the delta and delta-delta energy features[9].

The model set used in the connected digit recognition consists of 276 model units, including 264 inter-word context dependent model units for word junction acoustic events. Each inter-word context dependent unit is modeled by a 3-state continuous density HMM with 32 mixture for each state, except the body units which are modeled by a 4-state HMM with the same number of mixtures. The TI connected digit data corpus consists of 8623 sentence for training and 8700 sentence for testing. It has a vocabulary of 11 words $\{oh, 0, 1, \cdots, 9\}$. The length of the digit string is between 2 to 7.

The initial model is obtained from the conventional maximum likelihood approach. The performance of the model after maximum likelihood training is already very high. The string error rate of the initial ML model with unknown length decoding is 0.97%(84) and word error rate is 0.33%(93). The string model based minimum error rate training was applied to the ML trained model. The mathematical formulation of string model based minimum error rate training can be applied to the inter-word context dependent model units directly, once the string models needed for minimum string error rate training are identified. There is no grammar constraint on the competing strings and, therefore, unseen word contexts will occur in the competing string models.

Table 1. illustrates the performance comparison between the original model obtained from ML training and the model obtained from the string model based minimum error rate training. Using only one gender-independent model obtained from the string model based minimum error rate training, we achieved a string error rate of 0.72% with a corresponding word error rate of 0.24% with unknown length decoding. This is the best performance reported so far on this database. Comparing with the original model obtained from ML approach, the string and word error rate reduction obtained through minimum string error rate training is over 25%.

## 5. SUMMARY

In this paper, we studied the issues related to string level acoustic modeling in continuous speech recognition. The inter-word context dependent acoustic model units are modeled at the global string level based on the string level acoustic model matching. Experimental results indicate that a significant error rate reduction can be achieved through the proposed approach for models using inter-word context dependent model units. Under the unknown length decoding, the best performance obtained by a gender-independent model on the TI connected digit corpus is 0.24% word error rate and 0.72% string error rate.

## REFERENCES


[1] L.R. Bahl, P.F. Brown, P.V. De Souza and R.L. Mercer, "A New Algorithm for the estimation of hidden markov model parameters", *Proc. ICASSP'88.*

[2] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error rate training", IEEE Trans. on Signal Processing, Vol. 40, pp. 3043–3054 (1992).

[3] P.-C. Chang and B.-H. Juang "Discriminative template training for dynamic programming speech recognition", *Proc. ICASSP'92* pp493–496.

[4] W. Chou, C.-H. Lee and B.-H. Juang, "Minimum error rate training of hidden Markov models based on the $N$-best string models", *Proc. ICASSP'93*, Vol. 1, pp. 652–655.

[5] S. Katagiri, C.-H. Lee and B.-H. Juang, "Discriminative multi-layer feed-forward networks", *Neural Networks for Signal Processing* Proc. IEEE-SP workshop pp. 11–20, Princeton, Sept 1991.

[6] W. Chou, B.-H. Juang and C.-H. Lee, "Segmental GPD training of a HMM based speech recognizer", *Proc. ICASSP'92* pp. 473–476, 1991.

[7] W. Chou, T. Matsuoka, B.-H. Juang and C.-H. Lee, "An algorithm of high resolution and efficient multiple string hypothesization for continuous speech recognition using inter-word models", to appear in *Proc. ICASSP'94*, March 1994.

[8] C.-H. Lee et al, "Context dependent acoustic modeling for connected digit recognition", 1993 ASA Fall meeting, Denver, Oct 93.

[9] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini and A.E. Rosenberg, "Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition", *Computer Speech and Language*, pp. 103-127, 1992.