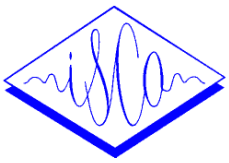

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SIGNAL CONDITIONED MINIMUM ERROR RATE TRAINING

Wu Chou, Mazin G. Rahim and Eric Buhrke

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

ABSTRACT

In this paper, a new approach, *signal conditioned minimum error training*, is proposed, where signal conditioning and minimum string error rate training are integrated into one process. The signal conditioning in this approach is based on hierarchical signal bias removal (HSBR), a novel extension of the signal bias removal algorithm. The HSBR is applied in conjunction with minimum string error rate training. In contrast to using a fixed codebook, the HSBR codebook used in our approach is derived from HMM parameters and updated with the HMMs during the process of minimum error rate training. As such, both HSBR signal conditioning and string model based minimum error rate training are based on the same set of HMMs. Experiments are performed on a connected digit database collected from the telephone network. The database covers various analog and digital network channels, different regional areas and a variety of telephone handsets. It is found that the proposed approach of signal conditioned minimum error rate training can lead to a significant reduction in recognition error rate. Based on a sub-word model consisting of various inter-word and intra-word context dependent (head-body-tail) model units, a 47% word error rate reduction is obtained through the proposed approach comparing with the model obtained from the conventional maximum likelihood (ML) training. This corresponds to an additional 27% word error rate reduction comparing with the inter-word context dependent sub-word model obtained from minimum error rate training where signal conditioning is not incorporated.

10.21437/Eurospeech.1995-132

1. Introduction

In this paper, an approach of *signal conditioned minimum error rate training* is proposed for the purpose of improving the speech recognition performance in adverse ambient conditions. The acoustic modeling in this study is based on a sub-word model consisting of various inter-word and intra-word context dependent model units. Words in the vocabulary are modeled by three types of sub-word model units, namely head units, body units and tail units. The head and tail units are inter-word context dependent model units. These units are acoustic driven defined by the joint acoustic events at the word junctions. Therefore, coarticulation in continuous speech is modeled explicitly, and a separate linguistic based lexicon describing the phonetic pronunciation of words in the vocabulary becomes unnecessary[1]. We refer to such a detailed acoustic model as a "head-body-tail" model.

The signal conditioning algorithm in the proposed approach is based on the hierarchical signal bias removal (HSBR), a new variant of SBR, and integrated with string model based minimum error rate training. The codebook of the HSBR in this approach is extracted from the HMM parameters and updated with the model during the process of minimum error rate training.

Experimental results of this study indicate that this approach can lead to a further performance improvement for models obtained in minimum error rate training. Comparing with the conventional maximum likelihood (ML) based approach, a word error rate reduction of 47% is obtained through the proposed approach. This corresponds to an additional 27% word error rate reduction comparing with the head-body-tail model obtained from minimum error rate training where signal conditioning is not incorporated. The performance improvements are from the enhanced acoustic modeling and signal conditioning without any increase in model size. The word error rate with unknown length decoding is 1.15% which is averaged over all the testing sets, including independent test sets collected from other independent efforts where their samples are not available for training.

2. String Model Based Minimum Error Rate Training

String model based minimum error rate training [2] is an approach based on the principle of string error rate minimization. The objective of minimizing the string error rate is embedded in a specially designed loss function. The problem of the "optimal classifier design" becomes that of finding the right parameter set of the discriminant function to minimize the "sample risk" defined as the average cost incurred in classifying the set of design samples in the training set. The string level acoustic variations are accommodated in this approach by modeling the basic speech recognition units at the whole utterance level.

The mathematical formulation of string model based minimum error training is described in [2] [4] and these definitions are listed below:

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

$$g(O, S_k, \Lambda) = \log f(O, \Theta_{S_k}, S_k \mid \Lambda), \qquad (1)$$

where S_k is the k-th best string, Λ is the HMM set used in the N-best decoding, Θ_k is the optimal path (state sequence) of the k-th string given the model set Λ, and $\log f(O, \Theta_{S_k}, S_k \mid \Lambda)$ is the related log-likelihood score on the optimal path of the k-th string.

For the lexical string S_{lex} from the training set, the discriminant function is given by

$$g(O, S_{lex}, \Lambda) = \log f(O, \Theta_{S_{lex}}, S_{lex} \mid \Lambda), \quad (2)$$

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

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

$$d(O, \Lambda) = -g(O, S_{lex}, \Lambda) + \log\{\frac{1}{N-1} \sum_{S_k \neq S_{lex}} e^{g(O, S_k, \Lambda)\eta}\}^{\frac{1}{\eta}}, \quad (3)$$

which provides an acoustic confusability measure between the correct and competing string models.

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

$$l(O, \Lambda) = \frac{1}{1 + e^{-\gamma d(O, \Lambda)}}, \quad (4)$$

where γ is a positive constant, which controls the slope of the sigmoid function.

(4) The model parameters are updated sequentially according to the generalized probabilistic descent (GPD) algorithm such that

$$\Lambda_{n+1} = \Lambda_n - \epsilon_n U_n \nabla l(O, \Lambda), \quad (5)$$

where ϵ_n is a sequence of step size parameters, and U_n is a sequence of positive definite matrices[4].

One important issue in continuous speech recognition is to improve string coverage during the training process, so that each model unit can be trained properly. However, unlike the words in the vocabulary, string coverage in the training material is always limited. In some cases, training materials are collected by using a very restrictive grammar during the process of data collection, and the perplexity in training data is unsuitably low. While in testing, the test speech utterances may have a much higher perplexity than the training data. This mismatch between training and testing can result in degradations in recognition performance. In string model based minimum error rate training, this problem is alleviated by incorporating N-best competing string models. The N-best string models are generated with a very loose constraint or without any grammar constraint. In this way, the competing string models are driven by acoustics. The training grammar is typically over generating to allow acoustically confusable strings to be included in training. Consequently, the string coverage as well as the coverage on the triphone and word contexts are significantly improved.

3. Hierarchical Signal Bias Removal (HSBR)

The hierarchical signal bias removal (HSBR) method [6] is a blind equalization scheme. It aims at reducing the acoustic mismatch between the training and various testing conditions. It combines the signal bias removal (SBR) approach [5] with hierarchical clustering method [8] where the size of the codebook is dynamically expanded for signal bias compensation.

In order to apply multiple biases in a one-pass decoding scenario, a "fuzzy" distortion criterion, similar to the one used in hierarchical spectral clustering [8, 9], is incorporated in the fundamental scheme of signal bias removal. The mathematical framework for this scheme which will be referred to as fuzzy SBR is presented in [6]. Fuzzy SBR consists of the following steps:

1. An estimate of the bias is computed for each cluster, such that K biases are generated for a codebook of size K. The j^{th} cluster bias is defined as

$$b_j = \frac{1}{T_j} \sum_{k=1}^{T_j} (y_{t_j(k)} - \mu_j), \qquad 1 \leq j \leq K, \quad (6)$$

where $y_{t_j(k)}$ represent input frames which are nearest to the cluster centroid μ_j, and T_j is the number of frames which are classified to the j^th cluster.

2. For each frame y_t, a frame dependent bias, $\bar{b}_t$, is constructed such that

$$\bar{b}_t = \sum_{j=1}^{K} b_j \hat{\alpha}_{t(j)}, \quad (7)$$

where

$$\hat{\alpha}_{t(j)} = \frac{\alpha_{t(j)}}{\sum_{i=1}^{K} \alpha_{t(i)}}. \quad (8)$$

The cluster weighting, $\alpha_{t(j)}$, computed between the input signal and the cluster centroid is defined as

$$\alpha_{t(j)} = \left[(y_t - \mu_j)^a\right]^{-1}, \quad (9)$$

where a is set to 2 in the current study for a Euclidean distance. Upon computing the frame-based bias, $\bar{b}_t$, then the input signal y_t is conditioned by the following equation:

$$\tilde{x}_t = y_t - \bar{b}_t. \quad (10)$$

An important problem associated with SBR is the selection of a reference codebook of an appropriate size. Hierarchical SBR (HSBR) is an approach which incorporates hierarchical clustering [8] with fuzzy SBR. This process is initiated from a codebook with a single entry. After each step of codebook splitting, fuzzy SBR is applied. The equalized feature vector sequence is used as the input for the next step. The fuzzy SBR performed in the next step is according to a codebook whose size is doubled. This process continuous until a specified maximum value, K_{max} of the codebook size is reached. In our study, the typical maximum codebook size K_{max} is set to 16.

4. Integrating Minimum String Error Rate Training with HSBR

Signal conditioning is important not only for testing but also for training. Training data typically encompass data collected through different channels and communication medias. Therefore, the term "matched" channel condition is no longer well defined. The proposed approach of signal conditioned minimum error rate training is based on the hierarchical signal bias removal and is applied in conjunction with string model based minimum error rate training. It differs from other approaches in that the HSBR and string model based minimum error rate training are integrated into one process. SBR schemes, including the fuzzy SBR described above, are dependent on the very existence of a reference SBR codebook $\{\mu_i\}$ which characterizes the acoustic space under the "matched" condition. Typically, the codebook is generated from the VQ clustering of all acoustic feature vectors in the training data. The codebook used in our approach is generated by clustering all the mean vectors of the hidden Markov models used in recognition. In this approach, the SBR codebook for signal conditioning becomes a function of the HMM model parameters and an integrated part of the acoustic modeling for pattern matching in recognition. Moreover, it is more efficient than the process of generating a codebook from all the training data, since it only involves the clustering operation of a very small amount of mean vectors.

In signal conditioned minimum error rate training, each training utterance is signal conditioned by applying HSBR, prior to being used in string model based minimum error rate training. The HSBR codebook generated from the HMMs is updated with the HMMs during the process of string model based minimum error rate training. Therefore, the model based HSBR codebook takes into account of HMM model parameter variations during the process of training. The HMM model parameters are updated to model the basic speech recognition units in the acoustic space corresponding to the signal conditioned training data, and signal conditioning is determined by the VQ clustering of the mean parameters from HMMs which characterize the patterns of the basic speech recognition units. A diagram of this approach is illustrated in Fig. 1. One advantage of this approach is that it can adapt to a new environment quite easily even if the training data from that environment are sparse.

It should be noted that string model based minimum error rate training aims to improve acoustic modeling of the basic speech recognition units based on a given training database. Although string model based minimum error rate training is effective in reducing the error rate under adverse conditions, the error rate and the recognition performance are conditioned unavoidably upon the acoustic and channel variations. On the other hand, signal conditioning reduces the acoustic and channel variations existing in the data used

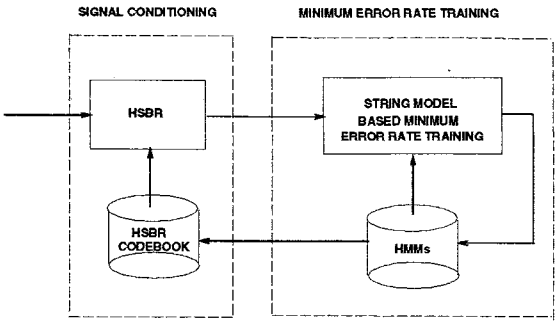

Figure 1: Diagram of signal conditioned minimum error rate training

in training and testing, making the acoustic space more homogeneous and less sensitive to ambient conditions. Based on signal conditioned training data, acoustic modeling is more focused, and consequently, the resolution of the model and the recognition performance can be improved further.

5. Experimental Results and Discussion

The experimental results are based on a connected digit database containing speech utterances recorded over the telephone network. It ranges in scope from one where talkers read prepared lists of digit strings, to one where voice inputs were actually used to access information. The database were collected over several different analog and digital network channels using a variety of telephone handsets. This amounted to 16087 and 21725 digit strings for training and testing, respectively.

Speech utterances were pre-processed using the following steps: pre-emphasis, blocking into frames of 30ms with a shift of 10ms, Hamming windowing, autocorrelation analysis, and finally performing LPC cepstral analysis which resulted in twelve littered cepstral coefficients. The twelve cepstral coefficients, the energy, and their first and second order derivatives were the input feature vectors for the HMM speech recognizer (see [7] for further details).

The recognition experiments were conducted in the following phases. In the first phase, we evaluated a system using context-dependent head-body-tail models. This configuration resulted in a grand total of 274 models, 831 states, and 6672 mixtures. The inter-word context dependent (head-body-tail) sub-word models was trained using the conventional maximum likelihood (ML) approach and achieved an average word error rates of 2.18% and an averaged string error rate of 8.9% (see Tables 1 and 2 under "Phase I").

In phase II of our study, the string model based minimum error rate training was applied to the sub-word models obtained in Phase I. Three iterations were performed over the entire training data. The average word error rate and string error rate were 1.58% and 6.2%, respectively (see Tables 1 and 2 under "Phase II"). Comparing with the inter-word context dependent sub-word model from phase I, this amounts to a

Method	Phase I	Phase II	Phase III	Phase IV
Wd_err (%)	2.18	1.58	1.35	1.15

Table 1: Word error rate for unknown length grammar.

Method	Phase I	Phase II	Phase III	Phase IV
Str_err (%)	8.9	6.2	5.6	4.8

Table 2: String error rate for unknown length grammar.

further reduction in the word error rate by 28% and string error rate by 30%. Note that in this phase the model size and the decoding strategy were identical as in phase I. What is different is the training procedure for generating inter-word context dependent model units.

In phase III, the HSBR method was applied only in testing based on the model obtained in Phase II. A codebook of size 16 was generated from the mean parameters of HMMs. Each testing utterance was signal conditioned prior to decoding. Tables 1 and 2 (under "Phase III") show that the average word error rate and string error rate when only using HSBR in testing with the model from phase II is 1.35% and 5.6%, respectively. These results represent a reduction in the word error rate by 15% and in the string error rate by 10% from phase II.

Finally, in phase IV, we evaluated a sub-word model obtained from the signal conditioned minimum string error rate training proposed in Section . Three iterations of signal conditioned minimum string error rate training were performed on the model in Phase II. An HSBR codebook of size 16 was generated from mean parameters of the HMMs used in recognition. During the signal conditioned minimum string error rate training, HSBR codebook was updated. The results in Tables 1 and 2 (under "Phase IV") show that signal conditioned minimum error training resulted in a further 20% word error rate reduction and a 14% string error rate reduction comparing with phase III. The final word error rate for the model obtained from the proposed approach of signal conditioned minimum error rate training obtained in phase IV is 1.15%, and the corresponding string error rate is 4.8% averaged over all test utterances in the database.

The results presented in Tables 1 and 2 demonstrate the efficacy of signal conditioned minimum error rate training for connected digit recognition. The overall reduction in the word error rate is 47% from the inter-word context dependent (head-body-tail) sub-word model obtained from conventional ML training (i.e., Phase I). This corresponds to an additional 27% word error rate reduction comparing with the inter-word context dependent sub-word model obtained from minimum string error rate training where signal conditioning is not incorporated (i.e., Phase II).

6. Summary

The results of this study indicate that a significant reduction in recognition error rate can be obtained by integrating a signal conditioning technique with minimum string error rate training. The improved speech recognition performance is made possible by both the new enhancements to minimum string error rate training algorithm permitting detailed training of the context dependent and position dependent acoustic model units in speech recognition and by the hierarchical signal bias removal algorithm presented in this paper. The integration of hierarchical signal bias removal with minimum string error rate training has been shown to be effective, resulting in error rate reduction over every region represented in the database. The interesting fact is that the performance improvement is quite uniform and on the matched conditions the performance is not compromised .

7. Acknowledgments

The authors acknowledge useful discussions with Biing-Hwang Juang and Chin-Hui Lee.

8. REFERENCES

1. Lee, C. H., Chou, W., Juang, B-H., Rabiner, L. R. and Wilpon, J. G., "Context-dependent acoustic modeling for connected digit recognition," *Proc. ASA*, Denver, CO, Oct.1993.
2. Chou, W., Lee, C-H., and Juang, B-H. "Minimum error rate training based on N-best string models", *Proc. ICASSP' 93* Vol II. pp. 652-655.
3. Chou, W., Lee, C-H., and Juang, B-H. "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition", *Proc. ICSLP' 94* pp. 439-442, Yokohama.
4. Chou, W., Juang, B-H., and Lee, C-H. "Segmental GPD Training of an Hidden Markov Model Based Speech Recognizer", *Proc. ICASSP92* pp. 473-476, 1992.
5. Rahim, M. and Juang, B-H. (1994) "Signal Bias Removal for Robust Telephone Speech Recognition in Adverse Environments," *Proc. ICASSP' 94*, **II.**
6. Rahim, M., Juang, B-H. Chou, W. and Buhrke, E., "Signal Conditioning Techniques for Robust Speech Recognition," submited to IEEE Signal Processing Letters.
7. 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.
8. Furui, S. (1986) "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. ASSP*, **34**, pp. 52-59.
9. Cung, H.M., and Normandin, Y. (1993) "Noise adaptation algorithms for robust speech recognition," *Speech Communication*, **12**, pp. 267-276.
10. Liu, F.-H., Stern, R., Acero, R., and Moreno, P. (1994) "Environment Normalization for Robust Speech Recognition using Direct Cepstral Normalization," *Proc. ICASSP' 94*, **II**, pp. 61-64.