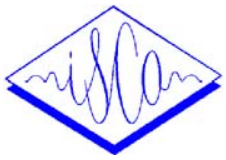

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

MAXIMUM-LIKELIHOOD UPDATES OF HMM DURATION PARAMETERS FOR DISCRIMINATIVE CONTINUOUS SPEECH RECOGNITION

Rathinavelu Chengalvarayan

Speech Processing Group, Lucent Speech Solutions
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

10.21437/ICSLP.1998-641

ABSTRACT

Previous studies showed that a significantly enhanced recognition performance can be achieved by incorporating information about HMM duration along with the cepstral parameters. The reestimation formula for the duration parameters have been derived in the past using fixed segmentation during K-means training and the duration statistics are always fixed throughout the additional minimum string error (MSE) training process. In this study, we update the duration parameters along with other model parameters during discriminative training iterations. The convergence property of the training property based on the MSE approach is investigated, and experimental results on wireline connected digit recognition task demonstrated a 6% word error rate reduction by using the newly trained duration model parameters as compared to fixed duartion parameters during MSE training.

1. INTRODUCTION

Explicit duration modeling has been shown to increase the effectiveness of hidden Markov models (HMM) in automatic speech recognition [1, 3, 10, 11, 14]. The HMM's to be discussed in this paper are based on continuous, mixture density models of the distribution of LPC derived parameter vectors [8]. In this study, we present results that demonstrate major improvements in our ability to recognize unconstrained strings of connected digits. We show that by incorporating new information about HMM duration along with the cepstral parameters, we can significantly enhance recognition performance. For scoring a given observation sequence using the internal duration model, a recursion of the Viterbi procedure is required. The recursion is considerably more costly than the implementation of the standard Viterbi scheme [8]. The post-processor durational model, on the other hand, uses the original Viterbi alignment procedure. Then for each word, the optimal model sequence is determined, and the duration of each model is obtained via a backtracking procedure. The loglikelihood is then augmented by the log duration probabilities (suitably weighted) to give the final score for the recognition decision, as shown

$$\log(S') = \log(S) + \alpha \sum_{j=1}^{N} \log(P_j(d_j)), \qquad (1)$$

where d_j is the number of frames occuring in model j, P_j is the duration probability of jth model, α is experimentally determined positive scaling constant and N is the total number of models in the model sequence of the given word. The model duration is modelled by a Gamma distributuion with

$$P(d) = \eta^{\nu} d^{(\nu-1)} \frac{exp(-\eta d)}{\gamma(\nu)}, \qquad (2)$$

with parameters ν and η and with means $\frac{\nu}{\eta}$ and varaiance $\frac{\nu}{\eta^2}$. The reestimation formula for η and ν have been derived in the past using fixed segmentation during K-means training and the duration statistics is always fixed throughout the additional MMI or discriminative or corrective training process [6, 8, 15]. The performance of speech recognition can be further improved by accurately modeling the duration of short speech events [10, 11]. In this study, we update the duration parameters along with other model parameters (mean, variance, mixture weights) during discriminative training iterations. The duration mean and standard deviation parameters are calculated using new segmentations which are obtained by using the current MSE-trained models for each given utterence in a sequential manner as exemplified in this study.

2. DISCRIMINATIVE MODEL PARAMETER ESTIMATION

We have used two methods for obtaining estimates of the HMM parameters namely the conventional MSE algorithm (HMM-I), and a more effective MSE training procedure which updates both the duration parameters along with other HMM parameters (HMM-II). The segmental K-means training procedure was used [7] to obtain an initial maximum likelihood estimation (MLE) based boot model for the subsequent discriminative training. The MSE training directly applies discriminative analysis techniques to string level acoustic model matching, thereby allowing minimum error rate training to be implemented at the string level [8]. A brief formulation of the MSE algorithm using generalized probabilistic descent (GPD) method is as follows:

- A discriminant function in MSE training is defined as

$$g(O, S_k, \Lambda) = \log f(O, \Theta_{S_k}, S_k|\Lambda),$$

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 state

sequence of the k-th string given the model set Λ, and $\log\ f(O, \Theta_{S_k}, S_k|\Lambda)$ is the related log-likelihood score on the optimal path of the k-th string.

- The misclassification measure is determined by

$$d(O,\Lambda) = -g(O,S_c,\Lambda) + \log\left(\frac{1}{N-1}\sum_{S_k \neq S_c} e^{g(O,S_k,\Lambda)}\right)$$

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

- The loss function is defined as

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

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

- The model parameters are updated sequentially according to the GPD algorithm

$$\Lambda_{n+1} \quad = \quad \Lambda_n - \epsilon \nabla l(O,\Lambda), \tag{3}$$

Λ_n is the parameter set at the nth iteration, $\nabla l(O,\Lambda)$ is the gradient of the loss function for the training sample O which belongs to the correct class c, and ϵ is a small positive learning constant.

During the model training phase, we call one complete pass through the training data set as an epoch. For the case of HMM-I based on string-by-string training, model parameters are updated several times over an epoch. Moreover the duration statistics are accumulated for every string and updated once over an epoch at the end of each MSE training iteration for HMM-II. Let $L(i,j)$ be the segment length for ith segment of jth HMM model, N_j be the total number of segments for jth model. Then the new mean and standard deviation estimates of the duration distribution can be determined for each model j as follows:

$$\mu_j \quad = \quad \frac{1}{N_j}\sum_{i=1}^{N_j} L(i,j) \tag{4}$$

$$\sigma_j \quad = \quad \frac{1}{N_j}\sum_{i=1}^{N_j} [L(i,j)]^2 - \mu_j^2. \tag{5}$$

Note that the HMM mean, variance and the mixture weights are updated sequentially for every given utterence where as the HMM duration parameters are updated once the entire training set is processed. And the combined sequential and batch updating process continues until the MSE training reaches the required number of iterations.

3. FRONT-END PROCESSING

The speech input is sampled at 8kHz and preemphasized using a first-order filter with a coefficient of 0.95. The samples are blocked into overlapping frames of 30 msec in duration, where the overlap is set to 20 msec. Each frame is windowed with a Hamming window and then processed using

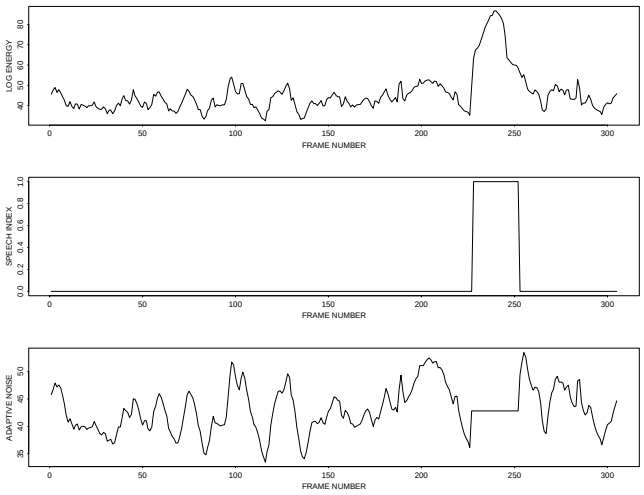

Figure 1. Typical energy measurement contours for the utterance "1". The top plot shows the original speech energy, middle plot shows the speech classification and the bottom plot provides the background adaptation.

a 10th-order LPC analyzer. The LPC coefficients are then converted to cepstral coefficients, where only the first 12 coefficients are retained. The basic recognizer feature set consists of 36 features that includes the 12 liftered cepstral coefficients and their first and second order derivatives [15]. Besides the cepstral based features, the normalized energy contour and its first and second order time derivatives are also computed. Thus, each speech frame becomes represented by a vector of 39 features. Note that the computation of all the higher order coefficients is performed over a segment of five frames [8, 9]. Since the signal has been recorded under various telephone conditions and with different transducer equipment, each cepstral vector was further processed using the two-level cepstral mean subtraction (2L-CMS) method in order to reduce the effect of channel distortion [2, 4, 12, 13]. The 2L-CMS technique is implemented in several steps [5]:

- Separate the frames of current utterance into two classes. If the current frame energy $E_t < \Theta + \gamma$ then the frame t belongs to class-I and Θ is updated using a leaky integrator $\Theta = \beta \times \Theta + (1-\beta)E_t$. Otherwise the frame belongs to class-II and β is the integration constant which determines the rate of convergence of the background energy estimate Θ.
- The background and the speech cepstral mean vectors are calculated for the whole uttereance.
- Finally the normalized cepstral features for each frame are computed by subtracting them by their respective cepstral means.

The above procedure is applied in both training and recognition [15]. To illustrate the nature of the signal classification, Figure 1 shows the actual frame energy trajectory and the corresponding speech index as well as the adapted back-

Databases	Training		Testing	
	Strings	Speakers	Strings	Speakers
DB1	2568	500	2649	500
DB2	2075	2075	1036	518
DB3	2639	2639	713	713
DB4	–	–	3063	200
DB5	–	–	4318	50
DB6	–	–	1335	1281
Total	7282	5214	13114	3262

Table 1. Regional distributions of spoken digit strings and the speaker population among the training and testing sets of the LSS_CD database.

ground energy trajectory for the isolated digit "1" spoken by a male speaker. It is observed that the 2L-CMS provides better speech and silence classification and further enhances the system performance.

4. SPEECH DATABASE

This section describes the database, LSS_CD, used in this study [15]. This database is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. These independent databases are denoted as DB1 through DB6. The LSS_CD database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over network channels using a variety of telephone handsets. Digit string lengths range from 1 to 16 digits. The LSS_CD database is divided into two sets: training and testing. The training set, DB1 through DB3, includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions. The testing set is designed to have data strings from both matched and mismatched environmental conditions and includes all six databases. All recordings in the training and testing set are valid digit strings, totaling 7282 and 13114 strings for training and testing, respectively. The data distribution of the training and testing set is shown in Table 1.

5. HMM ARCHITECTURE

Following feature analysis, each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models [8]. Each word in the vocabulary is divided into a head, a body, and a tail segment. To model inter-word coarticulation, each word consists of one body with multiple heads and multiple tails depending on the preceding and following contexts. In this paper, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Both the head and tail models are represented with 3 states, while the body models are represented with 4 states, each having 8 mixture components. Silence is modeled with a single state model having 32 mixture components. This configuration results in a total of 276 models, 837 states and 6720 mixture components. Training included updating all the parameters of the model, namely, means, variances, mixture gains and duration statistics using ML estimation followed by five epochs of MSE to further refine the estimate of the parameters. The number of competing string models was set to four and the step length was set to one during the model training phase. The length of the input digit strings are assumed to be unknown during both training and testing [15].

Type of Model	Training Scheme	
	ML Training	MSE Training
HMM-I	1.578%	0.986%
HMM-II	1.578%	0.927%

Table 2. Word error rate for an unknown-length grammar-based connected digit recognition task using the ML, conventional and newly proposed MSE trained models.

6. EXPERIMENTAL RESULTS

We have conducted experiments to verify the effectiveness of the proposed discriminative training process, using the continuous speech database, on the convergence property of the MSE training procedure and on wireline connected digit recognition performance. In Figure 2 we show empirical results on the behavior of the MSE training procedure for the *city name* continuous speech recognition task. The upper graph of Figure 2 shows the word error rates as a function of the epoch (a complete pass through the entire training data set is called an epoch) of the MSE training algorithm for the testing data. The solid lines are associated with MSE-trained conventional HMM (HMM-I), and the dotted lines with HMM generated using the new discriminative training (HMM-II). The lower graph of Figure 2 shows the average string loss for the entire training data set as a function of the training epoch. We observed that the recognition error rate monotonically decreases with the training epoch, and the average string loss monotonically decreases, both reaching their respective asymptotic values after five epoches of the training. The average loss decreases faster for the HMM-II than for the HMM-I, indicating the effectiveness of the newly introduced updated duration parameters. Similar characteristics in the recognition performance are also observed. This indicates that the original objective set out for minimizing the recognition error via the MSE training is accomplished and that the MSE training may be more effective for the HMM-II than the HMM-I.

The *connected digit* speech recognition results focusing on the comparative performances of the ML and MSE-trained HMM-II versus the HMM-I are summarized in Table 2. The results shown in Table 2 can be elaborated as follows. First, under all conditions the MSE training is superior to the ML training; the MSE-based recognizer achieves an average of 35% string error rate reduction, uniformly across all types

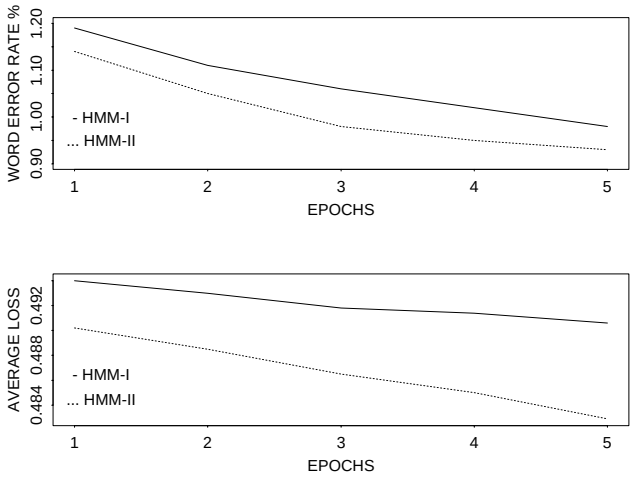

Figure 2. Convergence characteristics of the MSE training procedure. Top graph shows the word error rate for the "connected digit" recognition task and bottom graph shows the average string loss as a function of the training epoch.

of speech models. Second, for the ML-based recognizer, the HMM-II gives the same performance as compared with HMM-I since both the models have the same MLE trained boot model and duration parameters are updated during MSE training for HMM-II and in case of HMM-I the duration parameters remain the same throughout the MSE training. Thirdly, for the MSE-based recognizer, the HMM-II produces 0.92% word error rate which further yields about 6% reduction in word error rate compared with the HMM-I. Finally, the results presented in Table 2 demonstrate the efficacy of extended discriminative training procedure for *connected digit* continuous speech recognition.

7. CONCLUSIONS

In this work, the duration parameters have been updated along with other HMM parameters during discriminative training procedure. The duration mean and standard deviation parameters are estimated using the new Viterbi segmentation information which are further obtained by using the current MSE-trained models for each given utterence in a sequential manner as described in this study. This new approach is implemented and evaluated using modified MSE training methods. The convergence property of the training procedure based on the MSE approach is presented, which leads us to believe that the objective of minimizing the string error intended with the MSE criterion is achieved more effectively for the HMM-II than for the HMM-I. The experimental results on wireline connected digit recognition task yields a 6% word error rate reduction by using the newly trained duration model parameters as compared to fixed duartion parameters during MSE training. This suggests that the HMM duration parameters must also be updated during iterative discriminative training along with the mean, variance and mixture weight parameters.

REFERENCES

[1] A. Anastasakos, R. Schwartz, H. Shu, "Duration modeling in large vocabulary speech recognition", *Proc. ICASSP*, 1995, pp. 628-631.

[2] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *Journal of Acoustical Society of America*, Vol. 55, No. 6, 1974, pp. 1304-1312.

[3] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models", *Proc. ICASSP*, 1995, pp. 548-551.

[4] S. Furui, "Recent advances in robust speech recognition", *Proc. ESCA Workshop on Robust Speech REcognition*, Pont-a-Mousson, France, 1997.

[5] S.K. Gupta, F. Soong and R. Haimi-Cohen, "High accuracy connected digit recognition for mobile applications", *Proc. ICASSP*, 1996, pp. 57-60.

[6] B.H. Juang, L.R. Rabiner, S.E. Levinson and M.M. Sondhi, "Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition", *Proc. ICASSP*, 1985, pp. 9-12.

[7] B.H. Juang and L.R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 9, pp. 1639-1641, 1990.

[8] B.H. Juang, W. Chou and C.H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No.3, pp. 257-265, 1997.

[9] J.C. Junqua, D. Fohr, J.F. Mari, T.H. Applebaum and B.A. Hanson, "Time derivatives, cepstral normalization and spectral parameter filtering for continuously spelled names over the telephone", *Proc. EUROSPEECH*, 1995, pp. 1385-1388.

[10] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition", *Computer Speech and Language*, Vol. 1, 1986, pp. 29-45.

[11] C.D. Mitchell and L.H. Jamieson, "Modeling duration in a hidden Markov model with the exponential family", *Proc. ICASSP*, 1993, pp. 331-334.

[12] C. Mokbel, D. Jouvet and J. Monne, "Deconvolution of telephone line effects for speech recognition", *Speech Communication*, Vol. 19, No. 3, 1996, pp. 185-196.

[13] A. Rosenberg, C.H. Lee, F. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification", *Proc. ICSLP*, 1994, pp. 1835-1838.

[14] M. Russell and A.E. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition", *Proc. ICASSP*, 1987, pp. 2376-2379.

[15] D.L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features", *Proc. ICASSP*, 1998, pp. 21-24.