INFLUENCE OF OUTLIERS IN TRAINING THE PARAMETRIC TRAJECTORY MODELS FOR SPEECH RECOGNITION

Rathinavelu Chengalvarayan

Currently at: Speech Processing Group, Bell Labs
Lucent Technologies, Naperville, IL 60566, USA
Tel: (630) 224 6398, Fax: (630) 979 5915
Email: rathi@lucent.com

ABSTRACT

In this study, we developed a modified maximum likelihood (ML) algorithm for efficient computation in implemeting the minimum classification error (MCE) like training for optimally estimating the state-dependent polynomial coefficients in the trended HMM. We devised a new discriminative training method which controls the influence of outliers in the training data on the constructed models. The resulting models seem to provide correct recognition for confusable patterns. For alphabet recognition tasks, outlier emphasis resulted in improved performance. An error rate reduction of 14% is achieved for the linear trend and 7.5% is obtained for the constant trend models over the traditional ML training models.

1. INTRODUCTION

The formulation of the trended HMM (or trajectory-based HMM or nonstationary-state HMM) has been successfully used in automatic speech recognition applications for the past few years [3], [4], [5], [6], [7], [10]. The model parameters of the trended HMM (state-dependent time-varying means and variances) used in the past were trained using Viterbi-like algorithms based on the joint-state maximum likelihood principle (ML) [3]. The method of ML, however, need not be optimal in terms of minimizing classification error rate in recognition tasks in which the observation is assumed to be produced by one of the many source classes. Only the in-class information is available to train each model in ML approach, which leads to poor discriminative ability. Discrimination can be improved if out-of-class information is also used in training the models. Another alternative reestimation criterion, called minimum classification error (MCE) and maximum mutual information (MMI) training methods have been developed to improve the discriminating ability of ML criterion [2], [8], [9], [12]. This training approach takes into account other competing models and aims at minimizing the recognition error rate of the training data. The effectiveness of the MCE approach over the ML one demonstrated in our previous study for the trended HMM, however, is balanced by its significantly greater computation burden [11].

In this study, we developed a ML-like algorithm, which requires less computation, for efficient computation in implemeting the discriminative training for optimally estimating the state-dependent polynomial coefficients in the trended HMM. A new discriminative training method is proposed which controls the influence of outliers in the training data on the constructed models. The resulting models are shown to provide correct recognition for confusable patterns. The modeling proceedure is explained briefly: First ML models are constructed from the training data assuming the data is correct. Next, using these models the outliers are identified by recognizing the available training data. That is, the training tokens are weighted according to their relative match to their own word models and their degree of dissimilarity from the competing word models. This step is similar to calculating the sigmoid loss function, which approximates the classification error count [8]. That is, the loss function assigns near-zero penalty when an input is correctly classified and assigns a near-unity penalty when an input is misclassified. The outlier emphais can be done by setting the token weights equal to loss function and the outlier deemphasis is provided in the token weights by subtracting the loss function from one. And the ML re-estimation equations are adjusted to take into account the new set of weights for the training tokens. Finally one iteration of ML training is done by incorporating the new token weights, which weighs outliers either positively or negatively depending upon the task and applications.

2. MODEL PARAMETER ESTIMATION

The trended HMM is of a data-generative type and can be described as

$$\mathcal{O}_t \quad = \quad \sum_{p=0}^{P} \mathcal{B}_i(p)(t-\tau_i)^p + \mathcal{R}_t(\Sigma_i), \qquad (1)$$

where $\mathcal{O}_t$, $t = 1, 2, \cdots, T$ is a modeled observation data sequence of length T, within the HMM state indexed by i; $\mathcal{B}_i(p)$ are state-dependent polynomial regression coefficients of order P indexed by state i; and the term $\mathcal{R}_t$ is the stationary residual assumed to be independent and identically distributed (IID) and zero-mean Gaussian source characterized by state-dependent, but time-invariant covariance matrix Σ_i. The term $t-\tau_i$ represents the sojourn time in state i at time t, where τ_i registers the time when state i in the HMM is just entered before regression on time takes place. Each model state is characterized by a multivariate Gaussian density function with diagonal covariance matri-

ces in the form

$$b_i(\mathcal{O}_t|\tau_i) = \frac{(2\pi)^{\frac{-n}{2}}}{|\Sigma_i|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}\left[\mathcal{O}_t - \sum_{p=0}^{P} B_i(p)(t-\tau_i)^p\right]^{Tr} \Sigma_i^{-1}\left[\mathcal{O}_t - \sum_{p=0}^{P} B_i(p)(t-\tau_i)^p\right]\right) \quad (2)$$

where $B_i(p)$, Σ_i denote the polynomial means and variances of the i-th state of the model, $(t-\tau_i)$ is the sojourn time in state i at time t and n is the dimensionality. Superscripts Tr,-1 and the symbol $||$ denote the matrix transposition, inversion and determinant respectively. Based on the model j, the optimum state sequence $\Theta^j = \theta_1^j, \theta_2^j, \cdots, \theta_T^j$ for an input token $\mathcal{O} = \mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_T$ with T frames is obtained by means of Viterbi-algorithm [3]. Then, the log-likelihood is given by

$$g_j(\mathcal{O}, \Phi) = \sum_{t=1}^{T} \log b_{\theta_t^j}(\mathcal{O}_t|\tau_{\theta_t^j}) \quad (3)$$

Using an initial ML trained models each training token is assigned a weight determined by its degree of dissimilarity from the rest of the other training tokens.

2.1. Calculation of Training Token Weights

Let $g_j(\mathcal{O}, \Phi)$ denote the log-likelihood associated with the optimal state sequence Θ for the input token $\mathcal{O}$, obtained by applying the Viterbi algorithm using model Φ_j for the j-th class. Then, for the utterance $\mathcal{O}$ (from class c), the misclassification measure $d_c(\mathcal{O}, \Phi)$ is determined by

$$d_c(\mathcal{O}, \Phi) = -g_c(\mathcal{O}, \Phi) + g_\chi(\mathcal{O}, \Phi), \quad (4)$$

where χ denote the incorrect model with the highest log-likelihood (i.e., the most confusible class). In this definition, a negative value of $d_c(\mathcal{O}, \Phi)$ corresponds to a correct classification. The weights for the l-th training token of $j-th$ class in terms of misclassification measure can be calculated as

$$\mathcal{W}_{l,c} = \exp(-|d_c(\mathcal{O}, \Phi)|)$$

where $||$ denotes the absolute value. It can be seen from the above equation that a substantial token weighting is made when the absolute value of $d_c(\mathcal{O}, \Phi)$ is small — that is, when the training token is likely to be misclassified. On the other hand, when the absolute value of d_c is large, that is, when the input token is either unlikely to cause confusion or obviously an extreme outlier, then the amount of token weighting is accordingly reduced. However, a refined version of this index, defined by:

$$\mathcal{W}_{l,c} = 0.5 + \exp(-|d_c(\mathcal{O}, \Phi)| + \lambda)$$

seems more appropriate because the weights are more sensitive to outliers in the training data. The λ controls the influence of outliers in the above weighting expression. when λ tends to 1 the outlier is more likely emphasized, whereas an approximately -1 value for λ indicates an outlier deemphasis. Note that when all the token weights are set to one, we arrive at the traditional HMM training where all the tokens are given equal weighting in the parameter re-estimation process.

2.2. Estimation of Model Parameters

The segmentation step aims at finding a state sequence which maximizes the joint likelihood of observation sequence and state sequence. Once all the state boundaries are determined via the modified Viterbi segmentation step [3], learning the time-varying mean parameters in the trended HMM reduces essentially to the problem of polynomial regression. Here we present the general solution for the regression problem involving multiple observation tokens where each token can be a sub-sequence of a training utterence that has been segmented and assigned to a given state. In the remainder of this section, class index c will be omitted since in-class information is used in the ML training and hence each class model can be built independent of the other. Further, the new set of token weights are gracefully integrated into the parameter estimation formula as follows.

Let $\mathcal{O} = \{\mathcal{O}^1, \mathcal{O}^2, \cdots, \mathcal{O}^L\}$ denote a set of L feature vector sequences (i.e., L variable-length tokens), and let $\mathcal{O}^l = \{\mathcal{O}_1^l, \mathcal{O}_2^l, \cdots, \mathcal{O}_{T^l}^l\}$ denote the l-th sequence having the length of T^l frames. Define

$$\mathcal{X}_t(i) = [(t-\tau_i)^0 \ (t-\tau_i)^1 \ \cdots \ (t-\tau_i)^P]^{Tr}$$

is a $(P+1)\times 1$ vector of explanatory variables with $(t-\tau_i)$ representing the sojourn time in state i. The training token weighted maximum likelihood (ML-II) estimate for the polynomial coefficients becomes the solution to the regression equation

$$\mathcal{U}_i\,[\mathcal{B}_i(0)\ \mathcal{B}_i(1)\ \cdots\ \lambda\mathcal{B}_i(P)]^{Tr} = \mathcal{V}_i$$

where $\mathcal{U}_i$ and $\mathcal{V}_i$ are computed according to

$$\mathcal{U}_i = \frac{\sum_{l=1}^{L} \mathcal{W}_l \sum_{t=1}^{T^l} \gamma_t(i) \mathcal{X}_t(i)\,[\mathcal{X}_t(i)]^{Tr}}{\sum_{l=1}^{L}\sum_{t=1}^{T^l}\gamma_t(i)}$$

$$\mathcal{V}_i = \frac{\sum_{l=1}^{L} \mathcal{W}_l \sum_{t=1}^{T^l} \gamma_t(i) \mathcal{X}_t(i)\,\left[\mathcal{O}_t^l\right]^{Tr}}{\sum_{l=1}^{L}\sum_{t=1}^{T^l}\gamma_t(i)}$$

In the above equation, the quantity $\gamma_t(i)$ is set to be zero if the model stays in state i and is defined to be zero otherwise. The covariance matrix parameters are updated in a conventional maximum likelihood principle (ML-I) with uniform token weighting in the re-estimation procedure. Since the time-varying mean parameters are the most effective in modeling as well as in discriminating the different speech classes.

3. DATA FITTING ANALYSIS

The problem of speech classification can be viewed as a statistical data-fitting problem, where relative closeness in fitting an array of speech models to the unknown speech data sequence provides the classification decision. In order

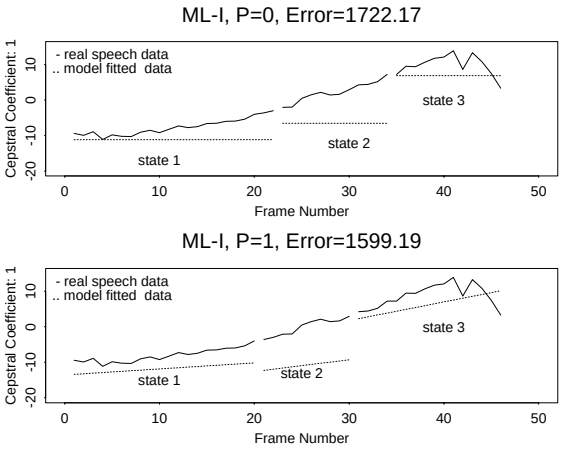

Figure 1. Fitting three-state ML-I trained **a** models to a speech data sequence.

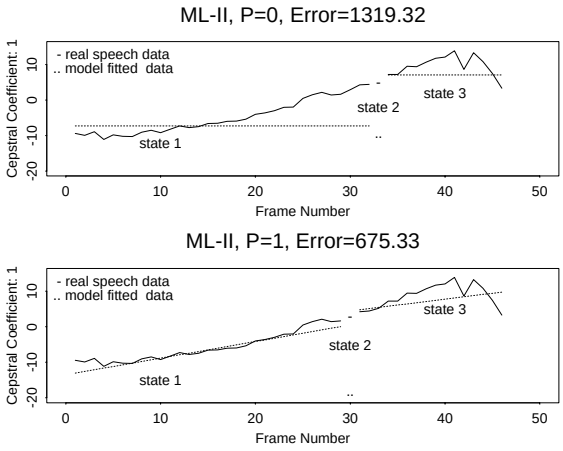

Figure 2. Fitting three-state ML-II trained **a** models to a speech data sequence.

to provide insights into the advantages of the ML-II training on the trended HMM, we report results of data-fitting experiments where both the conventional HMM and the trended HMM, trained with ML-I and with ML-II, respectively, are used to fit the acoustic observation data. Once the structure of the trended HMM is determined, the ML-II algorithm discussed in previous section is used to update the ML-trained trended HMM parameters by smoothly integrating the new set of outlier emphasized token weights into the re-estimation formula.

Figure 1 shows the results of fitting a test utterance (letter *a* from a first female speaker in the TI46 speech corpus) using the benchmark (P=0) and trended (P=1) HMMs. Use of first-order MFCC, C_1, as speech data here, shown in solid lines in Figure 1, is for illustration purposes only. Similar results are available for higher order cepstral coefficients. The two subplots of Figure 1 show the data-fitting results (dotted lines) for benchmark HMM (top) and trended HMM (bottom) when both models are trained by the conventional ML method. The two subplots of Figure 2 show the corresponding results (dotted lines) using the modified ML training with outlier emphasized trended HMMs. In all the plots, the solid lines are the real speech data, $\mathcal{O}_t$, of the C_1 sequence from a test token not used in updating the HMMs. The vertical axis represents the magnitude of C_1 and the horizontal time axis is expressed in terms of the frame number. For each sub-plot of Figure 1 and Figure 2, the two break-points in the otherwise continuous solid lines correspond to the frames at which the optimal state transitions occur from state one to state two, and from state two to state three, respectively. The dashed lines in all sub-plots of Figure 1 and Figure 2 are the four different trend functions, varying in the polynomial order ($P = 0$ or $P = 1$) and in the training procedure (ML-I or ML-II). These labels are shown at the head of each sub-plot, together with the data-fitting error computed by a linear summation of the residual squares over the states and over the state-bound time frames.

It is observed that the ML-II trained trended HMM fits the test token better than any other alternatives. For the benchmark HMM, error reduction in data fitting by incorporating the ML-II training goes from 1722 to 1319. The ML-II method for the linear trended HMM plays a more significant role of reducing the data-fitting error (a measure of better modeling capability) from 1599 to 675. This suggests that the time-varying mean parameters in the trended HMM are more effective in modeling the different speech patterns.

4. EXPERIMENTAL EVALUATION

The experiments conducted to evaluate the various trended HMMS are aimed at recognizing the 26 letters in the English alphabet, contained in the TI46 isolated word corpus. The training set consists of 10 tokens per word from two male and two female speakers (m1, m2, f1 and f2). The remaining 16 tokens per word for each of the above four speakers is used as test data. The preprocessor produces a vector of 13 Mel-frequency cepstral coefficients (MFCCs) for every 10 msec throughout the signal. The augmented feature vectors used for the trended HMM consist of 26-elements, with 13 cepstrum coefficients and 13 delta cepstra. The delta MFCCs are constructed by taking the difference between two frame forward and two frame backward of the MFCCs.

The main goal of the experiments designed in this study is to investigate the relative effectiveness of the token weighted training technique in comparison with the conventional maximum likelihood technique. Each word is represented by a single left-to-right, three-state HMM (no skips), with single Gaussian state observation densities. The covariance matrices in all the states of all the models are diagonal and are not tied. All transition probabilities are uniformly set to 0.5 (all transitions from a state are considered equally likely) and are not learned during the training process. The conventional trended HMM models (ML-I) are trained from training data using five-iterations of the ML training with single mixture for each state in the HMMs [3]. For the outlier emphasized approach, the initial model parameters are directly taken from the conventional HMM. The outlier emphasized (λ is set to 1) models (ML-II) are learned using an

Type of Model	Traditional ML Method (ML-I)	Weighted ML Method (ML-II)
P=0	80.11%	81.55%
P=1	83.59%	83.59%

Table 1. TI 26-alphabet classification rate using the conventional ML (left) and outlier emphasized ML (right) training methods

additional one iteration of the extended ML (training token weights are included in the parameter estimation process) algorithm as explained in the previous section.

Several sets of experiments are run to evaluate the alphabet classifiers constructed using two types of HMMs (stationary-state and trended) and two types of training (ML-I and ML-II) schemes. The overall performance of the alphabet classifiers, organized as the classification rate as a function of the polynomial trend function order ($P = 0$ for stationary-state HMMs and $P = 1$ for linearly trended HMMs) is summaried in Table 1 for the case of ML-I and ML-II training methods. The results shown in Table 1 can be elaborated as follows. First, under all conditions, the ML-II training is superior to the ML-I training. The ML-II based classifier achieves an error rate reduction of 7.5% for the constant trend models over the conventional trained models (ML-I). This error rate reduction is consistent with the previous related work. For example, with slightly different token weighting scheme using standard stationary-state HMM by others have reported an improvement of 9.8% in classification error rate to the problem of accent classification [1]. Second, for the ML-based classifier (Table 1), the trended HMM is superior to the stationary-state HMM, consistent with our earlier finding based on a different evaluation task [11]. Third, for the ML-II based classifier, superiority of the trended HMM over the stationary-state HMM becomes significantly greater than the ML case. Finally, the improvement in the classification rate in going from the ML-I to the ML-II training with use of the linear trended HMM is higher than that with the stationary-state HMM.

This shows that the behavior exhibited in Figures 1 and 2 in our data-fitting experiments is a dominant one, testifying to our conjecture that the ML-II training should be particularly effective for the trended HMM because of the greater degree of freedom existing in the modeled trajectories to allow for trajectory discrimination. The best result is achieved by using a combination of the trended HMM and the ML-II training algorithm, which produces an error rate reduction of 14% in moving from the ML-I training (83.59%) to the ML-II training (85.88%). We conclude from Table 1 that the outlier emphasized trended HMM trained by incorporating training token weights is superior to the conventional HMM.

5. CONCLUSIONS

In this study, we develop a ML-like algorithm for efficient computation in implemeting the discriminative training for optimally estimating the state-dependent polynomial coefficients in the trended HMM. A new discriminative training method is implemented which controls the influence of outliers in the training data on the constructed models. The derived learning technique is simple to implement and quite fast. An error rate reduction of 14% is achieved for the linear trend and 7.5% is achieved for the constant trend models over the traditional ML training models. The computation-related downside of the MCE like approach compared with the ML one has been slightly overcome by this approach.

REFERENCES

[1] L. M. Arslan, and J. H. L. Hansen, "Improved HMM Training and Scoring Strategies with Applications to Accent Classification", *Proc. ICASSP*, Vol. 2, pp. 589-592, Atlanta, May, 1996.

[2] W.Chow, C.H.Lee, B.H.Juang and F.K.Soong, "A Minimum Error Rate Pattern Recognition Approach to Speech Recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 8, No.1, 1994, pp. 5-31.

[3] L. Deng, M. Aksmanovic, D. Sun, and C. F. J. Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 507-520, October, 1994

[4] T. Fukada, Y. Sagisaka and K. Paliwal, "Model Parameter Estimation for Mixture Density Polynomial Segment Models", *Proc. ICASSP*, Vol. 2, pp. 1403-1406, Munich, April, 1997.

[5] O. Ghitza and M. Sondhi, "Hidden Markov Models with Templates as Non-Stationary States: An Application to Speech Recognition", *Computer Speech and Language*, Vol. 2, pp. 101-119, 1993.

[6] H. Gish and K. Ng, "Parametric Trajectory Models for Speech Recognition", *Proc. ICSLP*, Vol. 1, pp. 466-469, Philadelphia, October, 1996.

[7] W. J. Holmes and M. J. Russell, "Linear Trajectory Segmental HMMs", *IEEE Signal Processing Letters*, Vol. 4, No. 3, March, 1997.

[8] S. Katagiri, C. H. Lee, B. H. Juang, "New Discriminative Training Algorithms Based on the Generalized Probabilistic Descent Method", *IEEE Workshop on Neural Networks for Signal Processing*, pp. 299-309, Princeton, September, 1991.

[9] H. Li, J. P. Haton and Y. Gong, "On MMI Learning of Gaussian Mixture for Speaker Models", *Proc. EUROSPEECH*, Vol. 1, pp. 363-366, Madrid, September, 1995.

[10] M. Ostendorf, "From HMMs to Segment Models," in *Automatic Speech and Speaker Recognition – Advanced Topics,* C. Lee, F. Soong, and K. Paliwal (eds.), Kluwer Academic Publishers, pp. 185-210, 1996.

[11] C. Rathinavelu and L. Deng, "Speech Trajectory Discrimination Using the Minimum Classification Error Learning", *IEEE Transactions on Speech and Audio Processing*, in revision, 1997.

[12] V. Valtchev, J. J. Odell, P. C. Woodland and S. J. Young, "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition", *Proc. ICASSP*, Vol. 2, pp. 605-608, Atlanta, 1996.