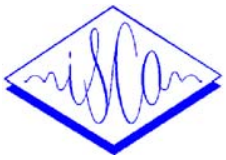

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PIECEWISE HMM DISCRIMINATIVE TRAINING

C. Chesta ⋆ and P. Laface ⋆ and M. Nigra ⋄

⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - I-10129 Torino, Italy e-mail chesta/laface@polito.it
⋄ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - I-10148 Torino, Italy e-mail mario.nigra@cselt.it

ABSTRACT

This paper address the problem of training HMMs using long files of uninterrupted speech with limited and constant memory requirements. The classical training algorithms usually require limited duration training utterances due to memory constraints for storing the generated trellis. Our solution allows to exploits databases that are transcribed, but not partitioned into sentences, using a sliding window Forward-Backward algorithm. This approach has been tested on the connected digits TI/NIST database and on long sequences of Italian digits. Our experimental results show that for a lookahead value L of about 1-2 sec it is possible to achieve reestimation counts that are affected by errors less than 1.e-7, producing similar reestimated models.

Another application of our sliding window Forward-Backward algorithm is MMIE training, that we have tested on the TI/NIST database connected digits using as a general model the recognition tree rather than the N-best hypotheses, or the word lattices.

1. INTRODUCTION

It is well known since long time that Viterbi decoding can be performed before the end of the observations because the optimal path is included among the best scoring paths at any given time (unless it is pruned by the beam search) and because all these paths are generated from a common ancestor (the "immortal node") [2].

In [1], on the basis of these observations, a piecewise Viterbi training algorithm has been proposed that requires a fixed amount of memory and can be used on unlimited length speech utterances.

In this paper we extend this approach to the EM reestimation performed through a sliding window Forward-Backward algorithm. Our approach does not require that a partial traceback locates an immortal node that identifies the unique survivor path to be passed to the next window. Instead, all the forward paths surviving a pruning operation with a large beam threshold are extended. The errors, therefore, do not depend on the estimated state alignment, but are related to the difference between the true state occupation probabilities (available only after all the observations have been processed) and the state occupation probabilities estimated looking ahead a short interval of time.

Another application of this approach is MMIE training [4], where the computation of the denominator of the objective function requires a substantial amount of memory resources for storing the trellis information.

2. APPROXIMATE FORWARD-BACKWARD

Let's consider the following approximate FB algorithm:

1. Perform the usual Forward algorithm, up to time L, storing the forward probability $\alpha_t(j)$ for each time frame t and state j of the estimated models
2. Perform the Backward algorithm from time L to time 0. The estimated backward probability $\tilde{\beta}_L(j)$ of each state j at time L is initialized as

$$\tilde{\beta}_L(j) = \frac{1}{\sum_s \alpha_L(s)} \tag{1}$$

so that the estimated probability of occupation of that state at time L is equal to its rescaled forward probability

$$\tilde{\gamma}_L^L(j) = \frac{\alpha_L(j)\tilde{\beta}_L(j)}{\sum_s \alpha_L(s)\tilde{\beta}_L(s)} = \frac{\alpha_L(j)}{\sum_s \alpha_L(s)} = \overline{\alpha}_L(j) \tag{2}$$

Of course, the Backward pass is initialized correctly for $L = T$, where T is the last frame of the utterance.

2.1. Error analysis

If the Backward pass is started before the utterance is completed ($L < T$) the estimated state occupation probabilities will be affected by an error.
In this Section we examine the behavior of these errors as a function of the distance between the frame L and the current time frame t.
Let's define b the state corresponding to the best forward path ($b = \text{argmax}_j(\alpha_L(j))$ and o the state with maximum occupation probability ($o = \text{argmax}_j(\gamma_L(j))$ respectively.
Typically, $\tilde{\gamma}_L^L(b)$ is an overestimate of $\gamma_L(b)$

$$\tilde{\gamma}_L^L(b) \geq \gamma_L(b) \tag{3}$$

because it relies only on the best forward path, and does not take into account the remaining observations $L+1, \ldots, T$. Moreover, since $\sum_s \gamma_L(s) = 1$ and $\sum_s \tilde{\gamma}_L^L(s) = 1$, and $\gamma_L(b)$ and $\tilde{\gamma}_L^L(o)$ are the most relevant contributions to their sums, $\tilde{\gamma}_L^L(o)$ is typically an underestimate of $\gamma_L(o)$.

$$\gamma_L(o) \geq \tilde{\gamma}_L^L(o) \tag{4}$$

This is definitely true if the estimated backward probability $\tilde{\beta}_L(b)$ is set to 1 rather than as defined in Eq. 1, while $\tilde{\beta}_L(j)$ for the remaining states $j \neq b$ is set to 0, as suggested in Fig. 1.

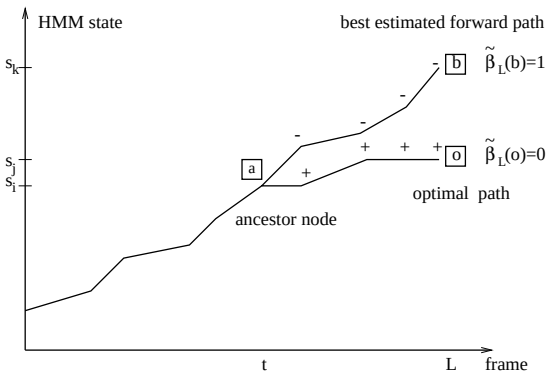

Figure 1: Best forward and optimal paths.

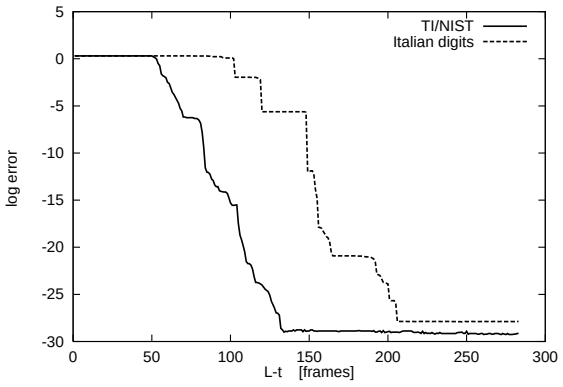

Figure 2: Maximum of the sum of state occupation probability errors (Γ_t^L) as a function of the distance $L-t$.

Our approximation of the true backward probabilities at time L, independently of the initialization of the backward probabilities at time L, produces *positive and negative errors* on the state occupation and backward probabilities at that time frame.
Since the backward probability is recursively computed as

$$\beta_{t-1}(j) = \sum_s \beta_t(s) \cdot b_t(s) \cdot a_{js} \tag{5}$$

the error on the estimated backward probability at time $t-1$

$$E^{\beta}_{t-1}(j) = \beta_{t-1}(j) - \tilde{\beta}_{t-1}(j) \tag{6}$$

will be back propagated as follows:

$$E^{\beta}_{t-1}(j) = \sum_s (\beta_t(s) - \tilde{\beta}_t(s)) \cdot b_t(s) \cdot a_{js} \tag{7}$$

if $\varepsilon^{\beta}_t(j)$ is the relative error on $\beta_t(j)$,

$$\varepsilon^{\beta}_t(j) = \frac{\beta_t(j) - \tilde{\beta}_t(j)}{\beta_t(j)} \tag{8}$$

we get

$$E^{\beta}_{t-1}(j) = \sum_s \varepsilon^{\beta}_t(s) \cdot \beta_t(s) \cdot b_t(s) \cdot a_{js} \tag{9}$$

and finally

$$\varepsilon^{\beta}_{t-1}(j) = \frac{\sum_s \varepsilon^{\beta}_t(s) \cdot \beta_t(s) \cdot b_t(s) \cdot a_{js}}{\sum_s \beta_t(s) \cdot b_t(s) \cdot a_{js}} \tag{10}$$

Since the relative error at each state, $\varepsilon^{\beta}_{t-1}(j)$, is computed as the the sum of the relative errors on its successor states, weighted by their *rescored* backward probability, the errors propagate along the best backward paths.
Since the best forward paths have a few common ancestors in the near past (a single common ancestor not too far in the past), the weighted sum of positive and negative errors along the best paths reduces the relative error that is propagated going back from the common ancestors, as illustrated in Fig. 1 where "+" and "-" marks correspond to positive and negative errors respectively.
States with very low probability of occupation may have high relative errors $\varepsilon^{\beta}_t(j)$, but their contribution to the reestimation counts is null because, for the sake of the efficiency, the reestimation counts are computed only for states with a probability of occupation greater than a given threshold.
The estimated state occupation probability $\tilde{\gamma}^L_t(j)$ can be expressed in terms of $\varepsilon^{\beta}_t(j)$ as

$$\tilde{\gamma}^L_t(j) = \frac{\alpha_t(j) \cdot \beta_t(j) \cdot (1 - \varepsilon^{\beta}_t(j))}{\sum_s \alpha_t(s) \cdot \beta_t(s) \cdot (1 - \varepsilon^{\beta}_t(s))} \tag{11}$$

and the error on the state occupation probability $E^{\gamma}_t(j)$ as

$$E^{\gamma}_t(j) = \frac{\alpha_t(j) \cdot \beta_t(j)}{\sum_s \alpha_t(s) \cdot \beta_t(s)} - \frac{\alpha_t(j) \cdot \beta_t(j) \cdot (1 - \varepsilon^{\beta}_t(j))}{\sum_s \alpha_t(s) \cdot \beta_t(s) \cdot (1 - \varepsilon^{\beta}_t(s))} \tag{12}$$

Going back from an ancestor state to another one along the best paths, the error $\varepsilon^{\beta}_t(j)$ decreases, $\tilde{\gamma}^L_t(j)$ approaches to the true value $\gamma_t(j)$, and the error $E^{\gamma}_t(j)$ approaches to zero.

2.2. Experimental results

To verify the behavior of the errors as a function of the distance between the Backward pass starting frame L and the current frame, the approximate FB algorithm has been used to compute the state occupation probabilities $\gamma_t(j) \;\; \forall j,t$ for 2156 adult male training utterances in the TI/NIST database.
The same experiment has been done on a connected digit telephone line Italian database, described in Section 5.3.
Setting the starting time of the Backward computation L to the last frame (T_u) of a given utterance (u) the true $\gamma_t(j) \forall t,j$ is computed, while if L assumes all the values in the interval $[1, T_u - 1]$ the estimated state occupation probabilities $\tilde{\gamma}^L_t(j) \;\; \forall t,j$ is obtained as a function of L. For each value of the distance between L and t we store the maximum of sum of the errors performed on the state occupation probabilities over all the states, i.e.

$$\Gamma^L_t = \max_{L-t} \sum_s \left(\gamma_t(s) - \tilde{\gamma}^L_t(s)\right)^2 \tag{13}$$

Fig. 2 shows how the errors on the state occupation probabilities, computed in step 2 of the approximate FB algorithm, decrease as a function of the distance $L-t$ (in 10ms frames). For the TI/NIST

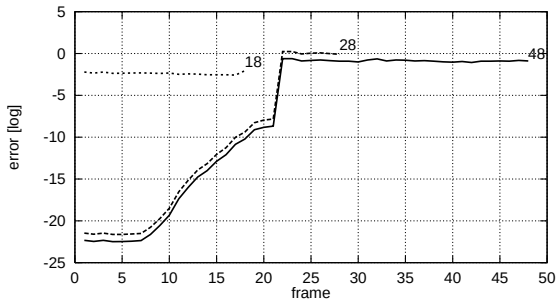

Figure 3: Γ_t^L as a function of 3 different starting frames L.

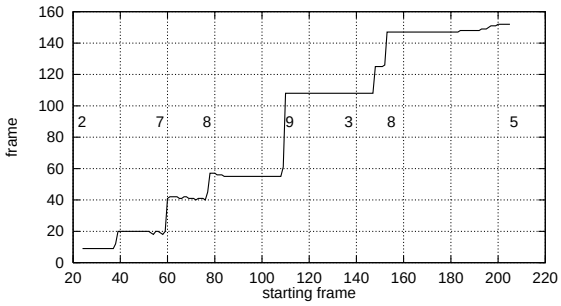

Figure 4: Rightmost frame for which $\Gamma_t^L) < 10^{-15}$ is obtained, as a function of the starting frame L.

database, in the worst case, after 150 (100) frames from the beginning of the Backward pass, the sum of the square errors of the state occupation probabilities is less than 10^{-30} (10^{-15}), that means that the reestimation counts are affected by an error less than 10^{-15} (10^{-7}).
For the more noisy, telephone speech, Italian database including long pauses, negligible errors are obtained, in the worst case, after 200 (150) frames from the beginning of the Backward pass.
A typical behavior of the errors Γ_t^L during the Backward pass, for a single short training utterance of $T = 50$ frames, with starting frame $L = 48, 28,$ and 18 respectively, is shown in Fig. 3. It can be noticed that the errors decrease to a negligible value even if the Backward pass starts from frame 28, while they cannot be reduced if the starting frame is not far enough (as shown for starting frame 18).
Fig. 4, that refers to the training utterance /2 7 8 9 3 8 5/ of the male speaker "ae_trm_c", plots, for every starting frame L of the Backward computation, the rightmost frame t for which a negligible Γ_t^L error (less than 10^{-15}) is obtained. Starting from frame $L = 140$, for example, a negligible error is obtained at frame $t = 108$. The labels in the figure mark the words boundaries in the utterance. The horizontal steps in the figure correspond to the frames where the best backward paths recombine into an "immortal node": these steps often corresponds to the word boundaries.

3. PIECEWISE FORWARD-BACKWARD

Since the errors on the state occupation probabilities $\gamma_t(j)$ become negligible at a relatively small distance from the beginning of the Backward pass (see Fig. 2) we can devise the following sliding window FB algorithm (with a lookahead of L frames):

1. Set the initial time $t = 0$
2. Perform the usual Forward algorithm, up to time L, obtaining and storing the forward probability for each state of the estimated models
3. Continue the Forward algorithm, up to time $t + 2L$
4. Perform the Backward algorithm from time $t + 2L$ to time t. The backward probability of each state at time $t + 2L$ is initialized according to Eq. 1.
   During the Backward step the reestimation counts are updated only for the trellis states included in the interval of time t to $t + L$, where the $E_t^\gamma(j)$ errors are negligible. It is worth noting that for this evaluation only two backward "trellis columns" are needed.
5. The memory used for storing the trellis information from time t to time $t + L - 1$ can be recovered.
6. Set the new window beginning time, $t = t + L$, and go to step 3

As shown in Section 2 given a large enough lookahead value L, these errors decrease going backward so that the sum of the deviations of the state occupation probabilities computed over all the times from 0 to $t + L$ and over all the states becomes very low.
For the experiments reported in Section 5.1 and 5.3, the lookahaed value L has been fixed to 100 and 160 respectively.

4. TREE BASED MMIE TRAINING

In MMIE training [4] for continuous speech of long sentences, the computation of the denominator of the objective function requires a substantial amount of memory resources for storing the trellis information.
Since our sliding window FB algorithm computes very good approximations of the state occupation probabilities, it can be used for MMIE training of very long sentences.
Another advantage offered by the piecewise discriminative training is the possibility to use the recognition tree as the general model, rather than the N-best hypotheses, or the word lattices. This avoids the assumption often made, for the sake of computation effectiveness, that the N-best hypotheses, or the word lattices do not change during training. Of course the computational load of this approach is greater than the standard ones, but it should be noticed that the restimation set for MMIE is always a subset of the training database. This approach has been tested of the TI/NIST database connected digits as reported in the next Section.

5. EXPERIMENTAL RESULTS

5.1. Piecewise Forward-Backward

Most of the experiments have been performed on the 20KHz TI/NIST connected digit corpus of adult speakers including 8700 sentence (28583 words) for testing. The signal is passed through a preemphasis filter and every 10 ms a 20 ms Hamming window is applied. A 512 point FFT is then performed and the frequency range up to 8 KHz subdivided into 20 Mel-scale filters is used to obtain 12 cepstral coefficients.
The observation vector paper includes 39 parameters: 12 liftered and high-pass filtered cepstral coefficients ($C_1 \div C_{12}$), and their

Training	Mixtures	Densities	sub/del/ins	WER (%)	SER (%)
MLE	1	1548	86/53/11	150 (0.52%)	133 (1.53%)
MMIE	1	1548	77/47/15	139 (0.49%)	122 (1.40%)
MLE	4	5480	50/19/8	77 (0.27%)	69 (0.79%)
MMIE	4	5480	49/25/11	85 (0.30%)	76 (0.87%)
MLE	8	9320	40/20/8	68 (0.24%)	62 (0.71%)
MMIE	8	9320	39/25/9	73 (0.25%)	66 (0.76%)

Table 1: Results for MLE and MMI trained models on the TI/NIST database

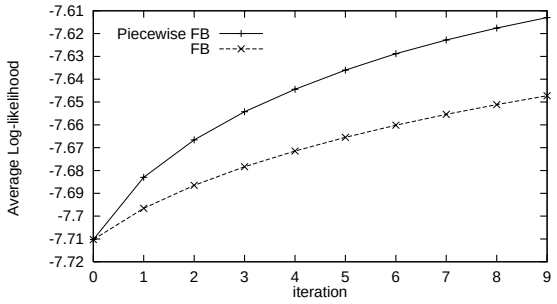

Figure 5: Average log-likelihood per frame after each iteration.

first and second order derivatives, the energy, and its first and second order derivatives.

The results shown in Table 1 have been obtained with unknown length decoding using two gender dependent acoustic models per word, one for short and the other for long duration utterances. The models were trained as introduced in [3], with a maximum of 1, 4 or 8 Gaussian densities per state and a single state silence model with 16 Gaussian densities.

Using the new sliding window FB algorithm for MLE training, we obtained *exactly the same* recognition results, in terms of word and string error rates, with respect to the models estimated using the classical FB. The average log-likelihood per frame after each iteration of the FB and of the sliding window FB algorithms for this training database are compared in Fig. 5. Surprisingly, the sliding window FB produces models that fit the training data better than the standard FB.

5.2. Piecewise Tree MMIE

As far as the tree based MMIE training is concerned, the results reported in Table 1 show that we were able to reduce the word and sentence error rate only for models with one Gaussian mixture per state, while only the substitution error rate has been slightly improved for more complex models, unfortunately with an increase of the insertion and deletion rates. Of course, it is difficult to improve the performance of accurate acoustic models trained with MLE, but the deluding performance of MMIE training on our special short and long models may be explained condidering that the tree based MMIE approach tries to discriminate between long and short models of the same word, probably reducing the robusteness of the competing models.

Training	sub/del/ins	WER (%)	SER (%)
MLE	179/112/65	356 (0.92%)	231 (9.3%)
MMIE	176/118/75	369 (0.96%)	236 (9.5%)

Table 2: Performance on the Italian database

5.3. Italian telephone digit database

A second experiment has been performed on a 8KHz sampled, telephone line, connected digit corpus including 8539 sentence for training and 2472 sentences (38533 words) for testing. The training sentences are composed of utterances including up to 16 digits, most of the test sentences are credit card numbers [3].
The same preprocessing is performed on the signal, but a 256 point FFT is applied to every 10ms window frame, and 12 Mel cepstral coefficients are computed. The energy and the high-pass filtered cepstral parameters and their first and second order derivatives are included in the observation frame.
The performance on the Italian database using long and short models with a maximum of 4 Gaussian per state are similar to the ones reported in the Section 5.2, and are detailed in Table 2.

6. CONCLUSIONS

The effectiveness of a piecewise FB training approach has been assessed comparing the results of MLE and Piecewise MLE training for the TI/NIST database, and for long sequences of Italian digits. This method has been also exploited for MMIE training with the recognition tree used as general model for the computation of the denominator of the MMIE objective function.

7. REFERENCES

[1] G. Boulianne, and al., “Books on Tape as Training Data for Continuous Speech Recognition”, Speech Communication, n.14, pp. 61-70, 1994.

[2] J.S. Bridle, and al. “An Algorithm for Connected Word Recognition”. Proc. ICASSP, 1982, pp. 899-902.

[3] C. Chesta, P. Laface, and F. Ravera. “Connected Digit Recognition Using Short and Long Duration Models”, Proceedings of Int. Conference on Acoustic Speech and Signal Processing, Vol.2, pages 557–560, Phoenix, USA, 1999.

[4] Normandin Y., “Maximum Mutual Information Estimation of Hidden Markov Models”, In “Automatic Speech and Speaker Recognition”, C.-H. Lee, F.K. Soong, K.K. Paliwal eds., pp. 57–81, Kluwer Academic Publisher, 1996.