

Speech Recognition with State Duration Constrained Maximum Likelihood HMMs

Kari Laurila, Markku Majaniemi, Ruikang Yang, Petri Haavisto

Nokia Research Center
Speech and Audio Systems Laboratory
P.O. Box 100, FIN-33721 Tampere, Finland
Email: kari.laurila@research.nokia.fi

ABSTRACT

In this paper we focus on the issue of duration modelling in the Hidden Markov Model (HMM) framework. We present a method to incorporate state duration hard limits into Maximum Likelihood (ML) training. In our new approach, only the state sequences fulfilling given state duration constraints, are used in the model parameter estimation. In addition, the same state duration constraints are used in the recognition phase. We also present a novel method to estimate the state duration hard limits. That is, the limits which maximise the likelihood of training observations given the overall (word level) duration constraints. Finally, recognition results using the obtained state duration constrained HMMs are compared to those of conventional HMMs showing the improvement.

1. INTRODUCTION

The speech recognition field can be divided into speaker-dependent and speaker-independent categories. Speaker-independent recognition is the more attractive approach from the users' point of view since no training is required, but speaker-dependent recognition can result in better recognition performance and enables user-specific commands. In many applications, like in voice-dialling, speaker-dependent models need to be utilised. Typically, the training of speaker-independent models has much less limitations than the training of speaker-dependent models. One of the most severe limitations in speaker-dependent model training is the small amount of training material. From the users' point of view, only one training pass would be desired. However, many training passes are typically required in order to create better models. Since it is important to design speech recognition systems that are attractive to the users we have focused on algorithm development required to make the former approach to perform well.

In single pass training many generally well performing approaches, like discriminative training, are not very usable. To be able to make a good speaker-dependent recogniser one must apply methods that are suitable considering the limitations in the training phase. Our approach is to bring logical restrictions to

the training and recognition processes. In speaker-dependent speech recognition more accurate duration modelling can be used than in speaker-independent speech recognition. This is natural since a single speaker tends to preserve his speaking style in similar situations (e.g. when speaking to a machine). Thus, a clever usage of duration modelling can be seen as a suitable means to improve the performance of speaker-dependent speech recognisers. Our main restriction is to force similar state segmentations in the training phase and in the recognition phase.

Duration modelling can also be utilised in speaker-independent speech recognition, although less severe restrictions than in speaker-dependent recognition can be applied due to the high variability in the speaking rates of different people.

2. THE PROPOSED ALGORITHM FOR STATE DURATION CONSTRAINTS

It is known that the conventional HMMs cannot model the temporal structures of speech effectively. It has been suggested by many researchers that explicit modelling of state durations can significantly increase the recognition rates [1-5]. State durations are usually modelled with certain distributions [1-4] and the probabilities produced by these distributions are added to the overall log-likelihood calculation. In the literature, bounded state durations used in the recognition phase have also been suggested [5].

In speaker-dependent recognition the models are often not very reliable (due to lack of training material) and the training occurs mostly in a reasonably silent environment (like in an office). Thus, if no restrictions are set for state-segmentations imposed by the Viterbi decoder then in a practical usage environment (e.g. a moving car in a highway) the state-segmentations can be quite different from the correct ones. Since wrong segmentations cause increased error rates it is in our interest to prevent them as effectively as we can. Our target is to force similar state segmentations in the training phase and in the recognition phase.

In the conventional ML training, the model parameters are estimated using all state sequences, most of which are unlikely to happen in the real cases but still contribute to the estimation. In another extreme case,

Viterbi training, only the best state sequence is used to estimate the model parameters which cannot result in robust estimates. In our new approach, only the state sequences that are likely to happen, and which fulfil given state duration constraints, are used in the estimation. In addition, the same state duration constraints are used in the recognition phase.

In the training phase we use state duration constrained ML training (SDML). We start initially with loose state duration bounds and gradually tighten the bounds.

2.1 State Duration Constraints within ML training

State duration bounds can be used in ML training by modifying existing forward and backward procedures. The modified forward procedure (assuming left-to-right models without skips) is given below:

$$\alpha_t(i) = P(\mathbf{o}_1 \mathbf{o}_2 \dots \mathbf{o}_t | q_t = i, \lambda, \varphi_i \leq d_i \leq \gamma_i) \text{ and}$$

$$\chi_t(i)_d = P(\mathbf{o}_1 \mathbf{o}_2 \dots \mathbf{o}_t | q_t = i, \lambda, \varphi_{i-1} \leq d_{i-1} \leq \gamma_{i-1}, d_i = d),$$

where q_t is the state at time t , φ_i and γ_i are the minimum and maximum durations of the state i , and d_i is the duration of the state i . The modified forward variable $\alpha_t(i)$ is the probability that $\mathbf{o}_1 \mathbf{o}_2 \dots \mathbf{o}_t$ are observed and that the state at time t is i on the condition that durations in states $\{1, 2, \dots, i-1, i\}$ are within the minimum and maximum limits. The additional forward variable $\chi_t(i)_d$ is the probability that $\mathbf{o}_1 \mathbf{o}_2 \dots \mathbf{o}_t$ are observed and that the state at time t is i on the condition that durations in states $\{1, 2, \dots, i-2, i-1\}$ are within the minimum and maximum limits and that the duration in state i is d .

1. Initialisation:

$$\chi_t(i)_d = 0, \text{ except } \chi_t(1)_1 = a_{0,1} b_1(\mathbf{o}_1), \text{ and}$$

$$\alpha_t(i) = 0, 1 \leq i \leq N, \text{ except } \alpha_t(1) = \chi_t(1)_1, \text{ if } \varphi_1 = 1$$

2. Induction

$$\chi_{t+1}(j)_1 = \alpha_t(j-1) a_{j-1,j} b_j(\mathbf{o}_{t+1}), 1 \leq j \leq N, \text{ and}$$

$$\chi_{t+1}(j)_m = \chi_t(j)_{m-1} a_{jj} b_j(\mathbf{o}_{t+1}), 2 \leq m \leq \gamma_j, 1 \leq j \leq N$$

$$\alpha_{t+1}(j) = \sum_{n=\varphi_j}^{\gamma_j} \chi_{t+1}(j)_n, 1 \leq j \leq N$$

3. Termination

$$P(\mathbf{O} | \lambda, \varphi, \gamma) = \alpha_T(N),$$

where T is the number of input frames.

In a similar way, a modified backward procedure can be defined.

2.2 Estimation of Optimal State Duration Constraints

Estimation of the state duration bounds is done using very loose beginning bounds (Notice that $\varphi_i=1$ and $\gamma_i=\infty$ correspond to the conventional ML training).

Duration bounds are then made more strict during ML training. I.e. the minimum state durations are estimated in the following way:

IF ($\sum_{s=1}^N \varphi_s < \text{MIN}$) **THEN**

FOR $i = 1$ **TO** N

{ $D_i = \varphi_i + 1$, and $D_j = \varphi_j$, where $j \neq i$
and $1 \leq j \leq N$,

$$P_i = P(\mathbf{O} | \lambda, D, \gamma) = \sum_{k=1}^T \alpha_{T_k}(N) \},$$

where MIN is the desired minimum word duration, T is the number of training tokens, T_k is the number of frames in the k 'th training token, and $\alpha_{T_k}(N)$ is the modified forward variable considering the state duration constraints D and γ evaluated at the T_k 'th (that is, the last) input frame in the state N .

$$m = \text{argmax}(P_i), \text{ where } 1 \leq i \leq N,$$

$$\varphi_m = \varphi_m + 1$$

Estimation of the maximum state durations can be done in a similar manner as the estimation of the minimum state durations.

2.3 Estimation of Word Level Duration Constraints

In one-pass speaker-dependent model training we use an end-point detection algorithm which utilises frame based power and zero-crossing rate values. A functional example of the algorithm is given in Fig 1. The black area is the time-domain power distribution of the utterance. The grey area corresponds to zero-crossing rate values. Four different thresholds are used. P_high and P_low for power, ZCR and ZCR_p for zero-crossing rate. Time-domain parameters ZCR_MaxDur , SIL_beg and SIL_end are also used. The end-point detection algorithm is the following:

A) Go forward until power of current frame exceeds P_high (1,3).

B) Come back until power of current frame is less than P_low (2,4).

C) Check if zero-crossing rate exceeds ZCR or zero-crossing rates in parallel frames exceed ZCR_p before current location (the search area is limited by SIL_beg or SIL_end). If so, come back to the first frame which

meets the requirements. In addition, if zero-crossing rate continues exceeding the threshold outside the search area, the end-point is moved to the last frame that still meets the threshold requirements (5).

Notice that SIL_end (see Fig 1.) is longer than SIL_beg because of possible unvoiced stops in the end of names.

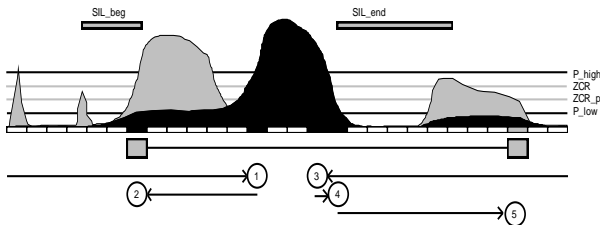


Fig. 1. A functional example of the end-point detection algorithm.

Word duration constraints are derived from the automatically end-pointed training utterances with the following rules:

$$\begin{aligned} \text{minimum name duration} &= \\ &0.8 * \text{training name duration,} \\ \text{maximum name duration} &= \\ &1.2 * \text{training name duration.} \end{aligned}$$

In speaker-independent model training we have manually checked labels for the beginnings and the endings of the words. The word duration constraints are set according to the duration statistics of the whole training database. We remove 3% of the shortest and the longest training utterances of each word and set the word duration constraints according to the remaining range. In Fig 2, a word duration histogram for the word “seven” is given. The corresponding minimum word duration is 0.29 seconds and the maximum word duration is 0.69 seconds.



Fig. 2. A word duration histogram for the word

“seven”.

3. RECOGNITION EXPERIMENTS

The proposed algorithm was first tested in a speaker-dependent isolated-word recognition task. The training of the models was done in an office environment using one training sample per model only. 30 Finnish first names including highly confusable ones were included in the vocabulary. The names were spoken by 5 native male speakers. 1-mixture variable-state HMMs (states were assigned to the models according to the lengths of the training utterances) for names and 1-mixture 3-state HMM for background noise were used in the experiments. A global variance vector was assigned for the name models.

In the test phase each name was said 3 times. In the highway noise test, car noise recorded in highway was added to the original files with approximately 0 dB SNR. In Table 1, the average performances of ML-training and SDML-training are presented. The usage of SDML-training reduced errors by 75% in highway noise.

	Office noise	Highway noise
ML training	96.9 %	70.0 %
SDML training	97.8 %	92.5 %

Table 1. ML training vs. SDML training in a speaker-dependent isolated-word recognition task.

The proposed algorithm was also tested in a speaker-independent connected-digit voice dialling task. In Table 2, the string level recognition rates achieved with connected-digit strings consisting of 6 digits and command “dial” are presented. The tests were carried out in parking place (motor off), city, and highway conditions in a car. With SDML multi-environment training over 29% reduction of string error rates was obtained over the results produced by the conventional ML training. In matched-environment training over 40% reduction was obtained. Given results are averages over different environments.

	Multi-environment	Matched environment
ML training	69 %	73 %
SDML training	78 %	84 %

Table 2. ML training vs. SDML training in a speaker-independent connected-word recognition task.

4. CONCLUSIONS

In this paper, a method to incorporate state duration hard limits into Maximum Likelihood (ML) training was presented. Also, a novel method to estimate the state duration hard limits was presented. State duration constraints applied in the training and the recognition phases were shown to significantly increase the recognition rates compared to the conventional ML training. In a speaker-dependent isolated-word recognition task errors decreased by 75% in noisy conditions due to SDML training.

REFERENCES

- [1] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol.77, No.2, pp. 257-286, 1989
 - [2] D. Burshtein, "Robust Parametric Modeling of Durations in Hidden Markov Models", Proc. ICASSP 1995, pp.548-551
 - [3] A. Anastasakos, R. Schwartz, H. Shu, "Duration Modeling in Large Vocabulary Speech Recognition", Proc. ICASSP 1995, pp.628-631
 - [4] S.V. Vaseghi, "State Duration Modelling in Hidden Markov Models", Signal Processing, Vol. 41, pp. 31-41, 1995
 - [5] H. Gu, C. Tseng, L. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations", IEEE Transactions on Signal Processing, Vol. 39, No. 8, 1991
- [1] L.R. Rabiner, "A Tutorial on Hidden Markov