# HMM TOPOLOGY SELECTION FOR ACCURATE ACOUSTIC AND DURATION MODELING

*C. Chesta ⋆ and P. Laface ⋆ and F. Ravera ⋄*

⋆ 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 franco.ravera@cselt.it

**ABSTRACT**

In this paper we show that accurate HMMs for connected word recognition can be obtained without context dependent modeling and discriminative training. To account for different speaking rates, we define two HMMs for each word that must be trained. The two models have the same, standard, left to right topology with the possibility of skipping one state, but each model has a different number of states, automatically selected.
Our simple modeling and training technique has been applied to connected digit recognition using the adult speaker portion of the TI/NIST corpus. The obtained results are comparable with the best ones reported in the literature for models with a larger number of densities.

## 1. INTRODUCTION

One of the main deficiency of the classical HMMs is related to inadequate modeling of the duration of the acoustic events associated with each state. Several solutions to this problem have been proposed. They rely on state duration modeling by means of discrete or continuous distributions that are more adequate to fit the temporal structure of speech. Another possibility is to use the state duration as an additional information for rescoring the hypotheses produced by Viterbi decoding in a post-processing approach. All these solutions, however, do not account for global spectral variations. Thus, they are not able to avoid recognition errors deriving by an incorrect time warping. Many errors often occur, indeed, because a sequence of observations is decoded by a few states - typically adsorbing low energy frames - with high probability and duration. The other states, instead, are rapidly traversed because their distributions do not fit well the remaining observations. These errors, therefore, do not depend on the intrinsic confusion of acoustically similar words, rather, the lack of good duration modeling and an incorrect time warping produces word hypotheses that are loosely related to the acoustics of the correct word.

In [5] we proposed an original approach to face these problems. It does not directly rely on state/word duration modeling, rather it models the global time variations of the spectral features of each word and their correlation in time: two important perceptual cues that are only partially exploited by standard HMMs. In particular, we rescore the probability produced by a conventional HMM system by means of the probability of a second very simple recognizer using word "temporal" models. The HMM system takes care of the local variations, while in the second system, the global time spectral variations of a word are modeled by means of two-dimensional cepstral features.
This post processing approach has given very good results for isolated word recognition [5]. Unfortunately, it produced marginal improvements only when we tried to rescore the N-best hypotheses produced by a connected digit recognizer. The reason of this behavior is that our approach heavily relies on correct alignments because the temporal models are trained on forced segmentations. In connected word recognition, instead, many errors are just due to incorrect alignments.

Another important issue for the classical HMMs is the so called trajectory folding phenomenon [4]. It happens because the characteristics of the speakers (their sex and speaking rate, for example) and all the other variabilities are merged into the models by using mixtures of densities associated to each state. This capability of merging highly variable information within a state, increasing the number of components of state mixtures, is one of the main reasons for the flexibility and the success of HMM modeling. This merging, however, has a cost in terms of discrimination capability: during recognition there is no mean to impose

| | Number of digits in sentence | | | | | |
|---|---|---|---|---|---|---|
| Position | 1 | 2 | 3 | 4 | 5 | 7 |
| 1 | 40 | 26 | 24 | 23 | 24 | 23 |
| 2 | | 43 | 26 | 29 | 29 | 27 |
| 3 | | | 39 | 28 | 32 | 49 |
| 4 | | | | 41 | 28 | 26 |
| 5 | | | | | 42 | 30 |
| 6 | | | | | | 26 |
| 7 | | | | | | 41 |

Table 1: Average duration in 10 ms frames of the utterances of digit ONE in the TI male speaker training set as a function of their position in the sentence

continuity constraints on the trajectory that a point in the parameter space follows as the articulatory system changes. Thus, an observation sequence can be recognized with high probability using a sequence of states and densities which have never been observed in the training set, leading to misrecognitions.
To solve these problems it has been proposed to train trajectory models [4] or trended HMM with state dependent, time varying Gaussian means [1].

In this work, we face the duration and trajectory folding problems in connected word recognition, with whole word models, by using a pragmatic approach that recognizes that some variability of the data is a priori known and can be modeled separately. The most evident source of variability is, of course, the female/male distinction, therefore, as usual in many systems, we train gender dependent models. Another important contribution to accurate modeling, however, is the definition of two HMMs for each word that must be trained: one "short" model for fast uttered words, and another "long" model for more articulated pronunciations. For short words like digits, the number of state of each model must be relatively large, in comparison with standard HMMs, so that it accounts for less than two frames per sentence on the average. We will show that, even if the resulting system has a relatively large number of states, good results are obtained with a reduced number of densities per state on the adult database of the TI/NIST connected digit corpus.
The organization of the paper is as follows. Section 2 introduces the motivations for different sets and topologies of models and illustrates the approach used to obtain automatically the number of states for each word model. Section 3 details the model training procedure. Finally, the results obtained using the set of models introduced in Section 2 are presented in Section 4.

## 2. MODEL TOPOLOGY SELECTION

As introduced in the previous Section, our simple approach toward accurate acoustic and duration modeling for whole word connected word recognition, defines a "short" and a "long", gender dependent, HMM for each word that must be trained. This solution tries also to reduce the trajectory folding problem.
The rational behind this choice is to account for different speaking rates, occurring not only in different utterances of the speakers, but also within a connected word utterance of the same speaker.
This effect is shown in Table 1. It presents the average duration - in frames of 10 ms - of the utterances of digit ONE in the TI/NIST training corpus of adult speakers, as a function of their position in the sentence. The sentences in the TI corpus include strings of length 1, 2, 3, 4, 5, and 7 quite uniformly distributed, thus, these distributions are similar for other digits.
The average duration of an utterance of digit ONE is 300ms, corresponding to 30 frames in our system. Looking at Table 1, it is interesting to note that:

- isolated words last more than average
- the duration of the first word is always less than the average duration and is pronounced faster than the other digits in the sentence
- the duration of the last word in the sentence is always greater than the average duration, and it is preceded by a short word pronunciation
- in the middle of long sentences there is a prepausal lengthening effect, clearly evident for sentences including 5 and 7 digits.

A single model, therefore, even if it is provided with skip transitions, don't seem adequate neither for duration nor for accurate acoustic modeling. The latter is true because the acoustic realizations of fast and slowly uttered words are likely to be different.
Our models have the same, standard, left to right topology, with the possibility of skipping one state, but each model has a different number of states.
For each word $w$, the number of states of its two HMMs is selected according to the following steps:

- The duration of every occurrence of word $w$ in the training set is generated by a forced alignment, using the set of models currently available.
- The histogram of the duration of all the ($N_w$) utterances of $w$ is obtained. Then the histogram values are cumulated up to $N_w/4$, $N_w/2$, and

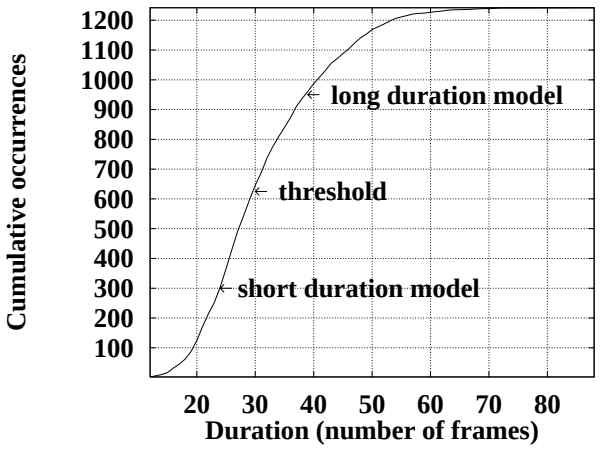


Figure 1: Cumulative distribution of the duration of the male speaker training utterances of digit ONE

$3/4 \cdot N_w$ respectively, and their corresponding duration values recorded.

- The number of states assigned to "short" and "long" duration models of word $w$ corresponds to the first and last duration value respectively. The central value is used, instead, as a duration threshold in the training procedure.

Figure 1 shows the cumulative distribution of the duration of the male speaker training utterances of digit ONE, and the number of states selected for this HMM model according to the above described procedure. Each word occurrence in the training set, then, contributes to the reestimation either of a "short" or of a "long" model: the decision is based on its duration compared with the duration threshold.
Table 2 shows the number of states obtained for each word model in the TI/NIST database. It is worth noting that the resulting the number of states of each model is comparable to the duration of its training samples, thus, the average occupation of each state is about one frame per sentence on the average. This contributes to the reduction of the trajectory folding phenomenon.

## 3. TRAINING

In our systems, training is performed by a few iteration of a segmental K-means Viterbi alignment procedure that allows the number of densities for each state to be automatically selected to fit the actual distribution of the training data as described in [3]. Since the number of states of each models corresponds to the average duration of short and long utterances of a word, it is large enough to allow accurate acoustic and duration modeling using a small number of densities per mixture. The maximum number of densities per state mixture was fixed to 8 for the reported experiments.
The bootstrap models are obtained as follows:

- Since isolated words have greater than average duration, it would be impossible to train reliable initial "short" models. Thus, both "short" and "long" isolated word bootstrap models are trained using all the isolated utterances in the training set.

- A segmental K-means Viterbi alignment is performed on the whole training set and the word duration statistics is collected

Training proceeds, then, through a few iterations of the following steps:

1. Generation, for each training sentence, of its HMM graph including the sequence of the appropriate "short" or "long" models according to the alignment obtained using the current set of models.

2. segmental K-means Viterbi alignment

Finally, several Baum-Welch estimation iterations are performed, keeping fixed the HMM graphs, until a convergence threshold is satisfied.

## 4. EXPERIMENTAL RESULTS

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 used in the recognition experiments reported in this paper includes *26 parameters only*: 12 liftered cepstral coefficients ($C_1 \div C_{12}$), 12 delta cepstral coefficients, the energy, and its first order derivative. Moreover, in these experiments, we did not perform any energy or cepstral mean normalization. The results shown in the Table 3, where the word and string error rates are reported, have been obtained with unknown length decoding using the following *gender dependent* acoustic models:

- The baseline system has a single model per digit with 8 Gaussian densities per state and a single state silence model with 16 Gaussian densities.

- The double model systems include two models per word with a maximum of 1, 4 or 8 Gaussian densities per state and a single state silence model with 16 Gaussian densities.

| Models | oh | zero | one | two | three | four | five | six | seven | eight | nine |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 16 | 34 | 22 | 22 | 22 | 28 | 30 | 24 | 40 | 20 | 20 |
| Short model | 20 | 35 | 24 | 20 | 24 | 27 | 30 | 33 | 36 | 20 | 28 |
| Long model | 35 | 49 | 39 | 34 | 38 | 43 | 50 | 52 | 47 | 31 | 43 |
| Threshold | 27 | 42 | 30 | 26 | 31 | 35 | 38 | 40 | 41 | 24 | 36 |

Table 2: Number of states for baseline, "short" and "long" duration HMMs, and duration threshold for the word models

| Acoustic models | No. of states | No. of densities | sub/del/ins | WER (%) | SER (%) |
|---|---|---|---|---|---|
| Baseline (8 G) | 278 | 4292 | 74/38/26 | 138 (0.58%) | 107 (1.23%) |
| Two models (1 G) | 1548 | 2244 | 106/75/20 | 201 (0.85%) | 172 (1.97%) |
| Two models (4 G) | 1518 | 5497 | 54/35/4 | 93 (0.39%) | 82 (0.94%) |
| Two models (8 G) | 1518 | 9021 | 52/31/7 | 90 (0.38%) | 79 (0.91%) |

Table 3: Performance comparison of the proposed modeling with respect to a classical gender dependent system

It is worth noting that, despite a very small word insertion penalty, the number of insertion errors is particularly low for the two models systems. This is due to the relatively large number of states used for the models, that cannot be easily traversed by observation sequences that do not fit well their distributions.
The obtained results are comparable with the best ones reported in the literature for models with a larger number of densities. In particular, the error rate of the 4 Gaussian double model system is comparable with the result in [2] - 93 (0.33%) WER 84 (0.97%) SER - for their MLE trained baseline system with 840 *context-dependent* states, 26880 Gaussian models, (they reach 0.24% WER and 0.72% SER with *discriminative training*), and with those presented in [6] - 99 (0.35%) WER 0.98% SER - using 716 states and 45824 densities, (their best result is 0.24% WER 0.74% SER using 22812 densities and *Linear Discriminant Analysis*).

## 5. CONCLUSIONS

In this paper we presented a simple modeling and training approach trying to cope with duration and trajectory folding problems.
The experimental results show that a significant error rate reduction can be obtained with respect to the classical HMM models. Moreover, our results are comparable with the best ones reported in the literature for models with a larger number of densities.
Since we did not use so far the second order derivatives, cepstral mean normalization, and discriminative training, we believe that good margins of improvement are still left for our system. The results of preliminary experiments using RASTA filtering and the second order derivatives in the observation vector are very promising and confirm our findings. We are also currently experimentig this approach for subword unit modeling.

## 6. REFERENCES


[1] C. Rathinavelu, and L. Deng. "The Trended HMM with Discriminatove Training for Phonetic Classification", Proc. International Conference on Spoken Language Processing, Philadelphia, PA, USA, pp. 1049–1052, 1996.

[2] W. Chou, C.-H. Lee, and B.-H Juang. "Minimum Error Rate Training of Inter-Word Context Dependent Acoustic Model Units in Speech Recognition", Proc. International Conference on Spoken Language Processing, Yokohama, Japan, pp. 439–442, 1994.

[3] L. Fissore, F. Ravera, and P. Laface, "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", Proc. EUROSPEECH 95, pp. 799–802, 1995.

[4] I. Irina, and H. Gong. "Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Model", Proceedings of Int. Conference on Acoustic Speech and Signal Processing, Vol.2,

[5] L. Fissore, P. Laface, and F. Ravera. "Using Word Temporal Structure in HMM Speech Recognition", Proceedings of Int. Conference on Acoustic Speech and Signal Processing, pp. 1395–1398, Munich, Germany, 1997.

[6] L. Welling, H. Ney, A. Eiden, and C. Forbrig. "Connected Digit Recognition Using Statistical Template Matching", In *Eurospeech 95*, pages 1483–1486, Madrid, 1995.