CONNECTED DIGIT RECOGNITION USING SHORT AND LONG DURATION MODELS

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. We train two HMMs for each word that 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. The two models account for different speaking rates that occur not only in different utterances of the speakers, but also within a connected word utterance of the same speaker.
This simple modeling technique has been applied to connected digit recognition using the adult speaker portion of the TI/NIST corpus giving the best results reported so far for this database. It has also been tested on telephone speech using long sequences of Italian digits (credit card numbers), giving better results with respect to classical models with a larger number of densities.

1. INTRODUCTION

Two important issues for the classical HMMs are duration modeling and the so called trajectory folding phenomenon [4]. The latter 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 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 [5] we have proposed to face the duration and trajectory folding problems by relying on the evidence 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 states 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. Although the resulting system has a relatively large number of states, similar or even better results are obtained with a reduced number of densities compared with standard models.

In this paper we report further improvements by using high-pass filtering of the cepstral parameters and their second order derivatives. This modeling technique has also been used for a telephone speech application using long sequences of Italian digits (credit card numbers). Even for this database better results are reported with respect to classical models with a larger number of densities.

The organization of the paper is as follows. Section 2 recalls 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. The model training procedure is described in Section 3 and the results obtained using the set of models introduced in Section 2 are presented in Section 4.1 and 4.2 respectively.

2. MODEL TOPOLOGY SELECTION

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.
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 [5].
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 $3/4 \cdot N_w$ 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 in the TI/NIST database, 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 1 shows the number of states obtained for each word model in the TI/NIST database and in the telephone speech Italian digit database described in Section 4.2
It is worth noting that the resulting 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.
It can be observed that the Italian digit models are longer than the corresponding English digits. This effect is partly due to the speaking style of some Italian speakers contributing to this database that did sometimes include long pauses between words, reducing their speaking rate.

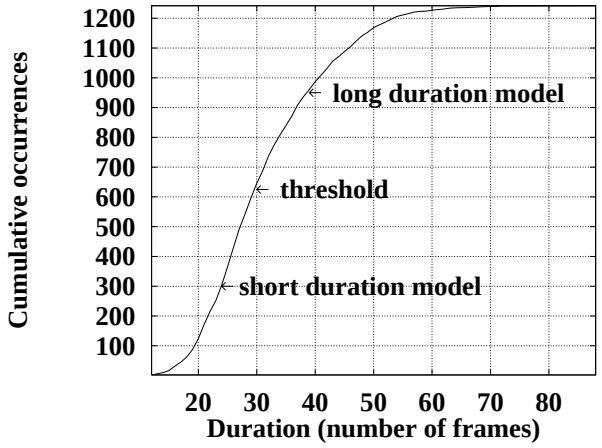

Figure 1: Cumulative distribution of the duration of the male speaker training utterances of digit ONE

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 and 4 for the reported experiments on the TI and Italian database respectively.
Training is performed by iterating 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

Few iterations are required to select the appropriate number of densities for each state, then, several Baum-Welch estimation iterations are performed, keeping fixed the HMM graphs, until a convergence threshold is satisfied.

4. EXPERIMENTAL RESULTS

4.1. TI/NIST database

The first set of experiments has been performed on the 20KHz TI/NIST connected digit corpus of adult speakers includ-

TI 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
Italian Models		zero	uno	due	tre	quattro	cinque	sei	sette	otto	nove
Baseline		36	32	28	26	40	40	30	38	32	30
Short model		38	32	32	24	40	40	34	42	36	30
Long model		58	48	52	40	62	64	56	68	56	50
Threshold		48	40	42	32	52	54	46	54	46	40

Table 1: Number of states for baseline, "short" and "long" duration HMMs, and duration thresholds

ing 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 up to 39 parameters: 12 liftered cepstral coefficients ($C_1 \div C_{12}$), and their first and second order derivatives, the energy, and its first and second order derivatives. In these experiments, we did not perform any energy normalization, but significant improvements were obtained by high-pass filtering the cepstral parameter.
The results in terms of word and string error rates are shown in the Table 2. They 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.

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 without high-pass filtering and second order derivatives 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*).
Using a maximum of 8 densities per state, and a total of 9320 densities, our best word and string error rate on the TI/NIST corpus are 0.24% and 0.71% respectively, that is the best performance reported so far on this database.

4.2. 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 (string length 16), but there are also several samples of 15 and 17 digit sequences.
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.
A set of *gender independent* double models per word has been trained to compare its performance with respect to an existing gender independent single model system. In particular:

- The baseline system has a single model per digit with a maximun of 16 Gaussian densities per state, a single state silence model with 32 densities, and a 26 state models for long pauses, with 16 Gaussian densities per state.
- The double model systems include two models per word with a maximum of 4 Gaussian densities per

	Acoustic models	Densities	sub/del/ins	WER (%)	SER (%)
No high-pass filtering No delta-delta	Baseline (8 G)	4292	74/38/26	138 (0.48%)	107 (1.23%)
	Two models (1 G)	1548	106/75/20	201 (0.70%)	172 (1.98%)
	Two models (4 G)	5497	54/35/4	93 (0.33%)	82 (0.94%)
	Two models (8 G)	9021	52/31/7	90 (0.31%)	79 (0.91%)
High-pass filtering No delta-delta	Baseline (8 G)	4292	55/29/23	107 (0.37%)	95 (1.09%)
	Two models (1 G)	1548	94/62/13	169 (0.59%)	147 (1.69%)
	Two models (4 G)	5212	52/27/8	87 (0.30%)	80 (0.92%)
	Two models (8 G)	9384	42/28/10	80 (0.28%)	72 (0.83%)
High-pass filtering Delta-delta	Baseline (8 G)	4292	52/26/23	101 (0.35%)	92 (1.06%)
	Two models (1 G)	1548	86/53/11	150 (0.52%)	133 (1.53%)
	Two models (4 G)	5480	50/19/8	77 (0.27%)	69 (0.79%)
	Two models (8 G)	9320	40/20/8	68 (0.24%)	62 (0.71%)

Table 2: Performance comparison of the proposed modeling on the TI/NIST database

	Acoustic models	Densities	sub/del/ins	WER (%)	SER (%)
High-pass filtering Delta-delta	Baseline (16 G)	5983	184/100/80	364 (0.94%)	238 (9.6%)
	Two models (4 G)	3623	179/112/65	356 (0.92%)	231 (9.3%)

Table 3: Performance comparison of the proposed modeling on the Italian database

state, a single state silence model with 16 Gaussian densities, and a 26 state model for long pauses, with 4 Gaussian densities per state.

The results of the comparison shown in the Table 3, confirm the effectiveness of our models even for a noisy telephone environment. Again, similar or slight better results have been obtained using a system with 60% of the densities of our baseline system. The relatively high sentence error rate, compared with a low word error rate, is not surprising because unknown length decoding is performed on very long digit strings.

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 or better than the best ones reported in the literature for models with a larger number of densities without requiring context dependent modeling and discriminative training. Due to the simplicity of this modeling, further improvements can be expected using LDA and discriminative training.
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, pp. 1395–1398, Munich, Germany, 1997.

[5] C. Chesta, P. Laface, and F. Ravera. “HMM Topology Selection for Accurate Acoustic and Duration Modeling”, Proc. International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[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.