CONTEXT-DEPENDENT DURATION MODELING

Daniel Willett

Temic Speech Dialog Systems, Ulm, Germany
daniel.willett@temic-sds.com

ABSTRACT

The paper proposes a context-dependent duration model and discusses its integration into a first-order hidden Markov model-based speech recognizer. The duration model allows the application of conditional duration probabilities that depend on the durations of neighboring HMM states. This way, it is capable of penalizing or fully preventing unusual durational relations of succeeding states. As the duration model is compiled into a dedicated first-order model topology it can be applied in a single-pass 1-best Viterbi decoder. In addition, this topology facilitates the integration of duration-dependent density functions. In experiments on connected digit recognition we see a relative word error reduction of about 16% with the proposed duration model and another 12% due to duration-dependent densities.

1. INTRODUCTION

In ordinary first-order HMM-based speech recognizers, the probability score $P_s(d)$ of staying in an HMM state s for exactly d frames, independent of the acoustic observation, is given by

$$P_s(d) = (1.0 - a_{ss}) \cdot a_{ss}^{d-1} \tag{1}$$

with a_{ss} representing the probability of a self-transition within s. Thus, the state duration probabilities are modeled as geometric distributions. It is well known, however, that these exponentially decreasing functions are rather poor models for duration probabilities. On the one hand, they are not capable of representing low probabilities for very short durations and on the other hand, they do not take any context durations or overall speaking rate into account but rather model each duration independently. Several studies have tried to overcome these obvious deficiencies by introducing more refined models for state or phone durations [1, 3, 4, 5, 6, 7, 8].

The basic approach of [6, 7, 8] replaces the geometric duration distributions with arbitrary probabilistic distribution functions f_s, such as Gaussian or gamma distributions. The overall score $\text{Score}_H(d_1, ...d_N)$ of a path through an HMM H of N states $s_1...s_N$ with durations $d_1...d_N$ is then expressed as

$$\text{Score}_H(d_1, ...d_N) = \prod_{i=1}^{N} f_{s_i}(d_i) \tag{2}$$

Such models are often referred to as Hidden Semi-Markov Models (HSMM). In case the distribution is discrete or in case it is mapped to a discrete probability distribution, the resulting higher-order model can be represented as a first-order model by expansion of each state into a chain of substates, as depicted in Fig. 1.

Often, first-order HMMs that represent higher-order HMMs through dedicated topologies are referred to as Expanded-State Hidden Markov Models (ESHMM). In an ESHMM, duration probabilities are modeled as discrete distributions and are stored in the transition probabilities. Earlier approaches [4] of applying minimum and maximum durations per HMM state are special cases of ESHMM. We will refer to a state of the higher-order model as well as to the set of substates that represent it as *superstate*.

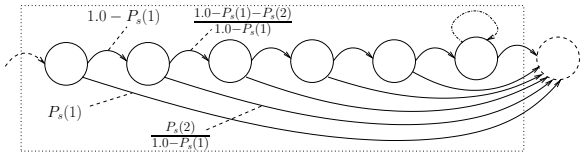

Fig. 1. A superstate of an Expanded-State Hidden Markov Model

The reported performance improvements with HSMM and ESHMM are only very small and rather disappointing. While providing a more accurate modeling of state durations, these models lack the ability to estimate duration probabilities conditioned on the neighboring states' durations or the overall speaking rate.

The approach in [1] tries to incorporate the correlation of durations within words by rescoring N-best lists according to trained word-based phone duration relations. The reported absolute improvement of 1% on Switchboard is notable, but the approach raises many questions concerning word clustering, smoothing and the treatment of unseen words. Povey [5] found it to be very hard to tune and gained only very limited improvements. Other approaches, such as the one in [3], try to make use of speaking rate estimates. Robust speaking rate estimates are best computed over entire utterances, so that this approach too requires an expensive N-best rescoring scheme. The measured marginal improvements do not quite justify this effort. Furthermore, these approaches are hardly applicable in online decoding scenarios.

Therefore, this study tries to develop a duration model that can be integrated into a first-order model-based decoder and trainer while at the same time allowing state duration probabilities to be conditioned on the duration of neighboring states in order to penalize unwanted durational behavior. As a side effect, we will see that the resulting model topology enables a straightforward extension to duration-dependent acoustic output distribution functions.

2. A BIGRAM EXPANDED STATE DURATION MODEL

The principal approach we propose extends the idea of state expansion by expanding each HMM state into an even larger automaton, capable of representing a durational bigram model. In this network of rows of linearly connected substates of increasing length each row of substates represents the corresponding superstate observed for a certain number of time frames (see Fig.2). This representation encodes the superstate duration into each of its substates. Hence, the transition probabilities are capable of representing exact probabilities $P_s(d_s \mid d_{s-1})$ of staying in a state s for d_s frames given that the duration of the previous state was d_{s-1} (see Fig.3). For now, we assume all substates of a superstate to share the same output distribution function (full tying of substates).

In the detailed scheme we are following, the maximum depth M_s up to which a state s is being expanded is either fixed (denoted later as $M_s = n$) or linearly dependent through some tuning parameter α on its average duration.

$$M_s = \lceil \alpha \cdot \text{AverageDuration}(s) \rceil = \lceil \alpha / (1 - a_{ss}) \rceil \tag{3}$$

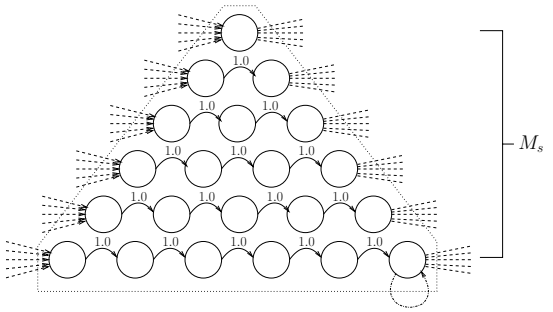

Fig. 2. Context-dependent ESHMM (CDESHMM) superstate

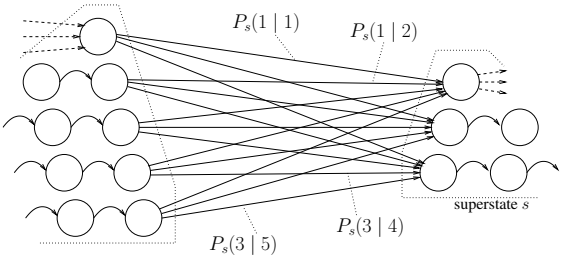

Fig. 3. CDESHMM superstate transitions (full connectivity)

A path through the N superstates of HMM H with durations $d_1, ...d_N$ is now scored with discrete transition probabilities conditioned on the preceeding durations according to Eq.4.

$$\text{Score}_H(d_1, ...d_N) = P_{s_1}(d_1) \prod_{i=2}^{N} P_{s_i}(d_i \mid d_{i-1}) \qquad (4)$$

Based on the assumption that neighboring durations are highly correlated, this is a more appropriate probabilistic model than the product over the unconditional duration probabilities of Eq.2.

In order to make sure they absorb exactly one feature frame, the substates do not permit self-transitions. Only the M_sth row is optionally given a loopable state (one with a self-transition) in order to enable this last row to also absorb observations of more than M_s frames. Certainly, the additional degree of freedom that comes along with the loopable substate in the last row also allows unwanted state durations and can contribute to recognition errors. The higher M_s and α, the less important these loopable substates become and fully avoiding loopable states becomes an option. Section 5 will show that even at a rather small α of 1.4, we found it to be beneficial in terms of recognition accuracy to fully sacrifice self-transitions, i.e. to limit the maximum duration.

3. PARAMETER REDUCTION, TYING AND SMOOTHING

In the CDESHMM scheme as proposed in the previous paragraph, a superstate s is expanded into M_s distinct substate paths of increasing length, with each path having distinct exit transition probabilities into the M_{s+1} entry states of the following superstate. This computes to $M_s \cdot M_{s+1}$ transition probabilities between each two superstates s and $s+1$. These transition probabilities need to be estimated robustly. The following paragraphs deal with the problem of reducing the number of transition probabilities for reasons of robustness and model size reduction.

3.1. Merging of substate rows of similar length

The merging of rows of similar length (similar duration) is an obvious procedure for reducing the number of transition parameters as well as the number of substates. In detail, what we experimented with is a scheme for merging all the rows of length 4 and 5, those of length 6,7 and 8, those of length 9,10,11 and 12, and so forth within all superstates s. It is depicted on the left hand side of Fig.4.

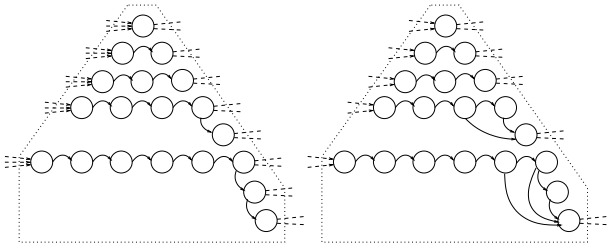

Fig. 4. Merged rows of similar duration (left), additional merging of exit states (right)

Expecially for large maximum depths M_s, this leads to a remarkable reduction of substates and transition probabilities. The transition probabilities in the exit states now represent the probability of the following duration being in some range from m to $\hat{m}$ given the exact duration of the current state. This can be regarded as a means of tying transition probabilities for the sake of robustness as well as model compression.

An even stronger tying of transition probabilities can be achieved by not only reducing the number of entry substates but by similarly reducing the number of exit substates. The right hand side of Fig.4 shows this scheme. The transitions between superstates now represent the probability of the following duration being in some range from m to $\hat{m}$ given that the current superstate duration is within some range n to $\hat{n}$. While further reducing the number of transition probabilities this procedure does not provide an additional reduction of required substates.

3.2. Reduced superstate connectivity

In the basic CDESHMM framework, decoding paths can proceed from an exit substate into each of the following superstate's entry substates. However, we have an idea (some prior knowledge) of which transition structure we want the model to learn. It is a smooth durational behavior with short durations following short durations and longer durations following longer durations.

Therefore, for the purpose of reduced superstate connectivity, we impose this topology by only allowing transitions between exit and entry substates, the duration of which (possibly each normalized by the average duration) is similar. Fig.5 shows the thinned out transition structure between two superstates, with the mean duration of the first being somewhat larger than that of the second.

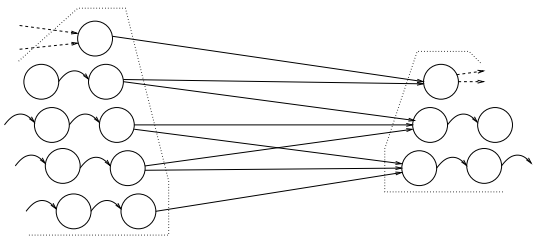

Fig. 5. Sparse superstate connectivity

Alternatively, we experimented with a more data-driven scheme. Similarly to an approach in [9] for mixture weight removal, we first permit all transitions and then gradually removed those that turned out to be taken with very low probability during parameter estimation. This, however, performed worse than the less expensive method described above that imposes a sparse transition structure from the start.

3.3. Transition smoothing

During HMM parameter estimation, probabilities $a_{s\hat{s}}$ of stepping from state s to state $\hat{s}$ are gained from dividing $\gamma_{s\hat{s}}$, the number of times the transition is being taken, by Γ_s, the total number of times (frames) that state s is being visited. For high numbers of $\gamma_{s\hat{s}}$ and Γ_s this is known to be a robust estimate. For lower numbers smoothing schemes such as those well known from statistical language modeling [2] can yield more robust estimates. In this study, we did not apply any smoothing of transition probabilities but instead put a focus on procedures as described above for effectively reducing the number of transition probabilities and the necessity for smoothing. This approach was motivated by the observation that strongly constraining the superstate transitions, as outlined in 3.2, turned out to be beneficial in terms of test data performance.

4. DURATION-DEPENDENT DENSITY FUNCTIONS

As stated before, the CDESHMM framework encodes the exact superstate duration into each of its substates. This way, it allows the straightforward integration of duration-dependent density functions for modeling the distribution of the acoustic observations. This could either be done implicitly by augmenting the acoustic feature vector with the current state duration or explicitely by relaxing the constraint of all substates of a superstate to share a single distribution function. For now, we evaluated different schemes for state clustering, as depicted in Fig. 6.

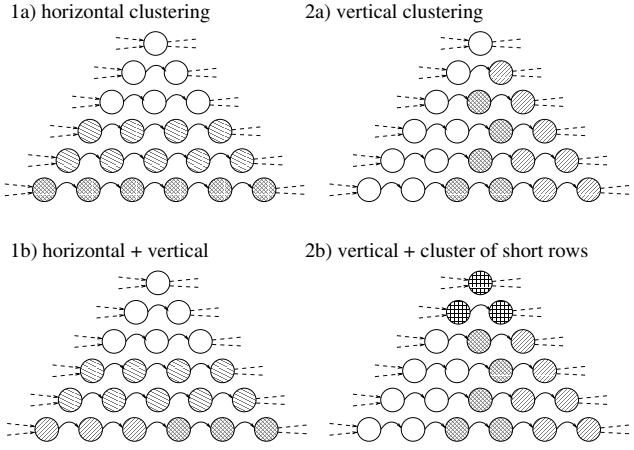

Fig. 6. Different types of substate clustering

In the approaches 1a and 2a, the substates of each superstate are divided into n_s classes where n_s depends on the amount of training data available for superstate s. In 1a, the clusters are oriented horizontally, resulting in different pdfs for different ranges of duration. In 2a, the state clustering is relaxed in the vertical direction, allowing a more accurate modeling of the temporal progression (trajectory) of the acoustic observation within a superstate. Scheme 1b starts off from the clustering in 1a and in addition relaxes the clustering of rows which receive a particularly large amount of training data. The approach 2b starts off from 2a and merges the rows of durations below n_s into a new dedicated cluster. The motivation is that with the number of substates being below n_s, these rows cannot make use of all introduced cluster pdfs and possibly rather harm the associated distributions than contribute to their robust estimation.

5. EXPERIMENT AND RESULTS

Experimental evaluations were carried out on whole word models for speaker-independent American English digit recognition. Training and test data consist of continuously spoken digit strings of varying length uttered in moving and stationary cars. Acoustic features are 11 MFCC coefficients transformed through LDA. The baseline HMMs have linear topology with 10 states each and tied-mixture probabilistic distribution functions.

Word error rates (WERs) achieved with CDESHMM in different configurations can be found in Table 1. Also, the table displays the baseline without state expansion and the best performance we managed to come up with in the context-independent ESHMM scheme (Fig.1). With a self-loop in each M_sth row, we see a small improvement from 3.55% down to a WER of 3.3% (at best) on the training data and similar relative improvements on the test data. When fully refraining from using loopable states, we see a very different and much more interesting trend. With an overly small maximum duration of $M_s = 5$ or $\alpha = 1.2$, we see a dramatic increase in WER on the training data, but still a small improvement on the test data. Even at the intermediate configuration of around $\alpha = 1.6$, we observe a marginal performance degradation on the training data but see a 14% relative word error reduction over the baseline on the test data. With a deep state expansion ($\alpha = 2.0$ or $M_s \geq 8$), we observe that removing self-loops is beneficial on both training and test data. Overall, refraining from having loopable states in the last row strongly constrains the models with a very positive effect on generalization. Hence, the following experiments do only make use of non-loopable substates.

configuration		with self-loops		without self-loops	
	av. # of substs.	training WER	test WER	training WER	test WER
baseline		3.55	**4.79**	-	-
ESHMM		3.51	4.71	3.23	4.49
fixed number of substate rows per state					
$M_s = 5$	15	3.34	4.52	5.52	4.32
$M_s = 6$	21	3.34	4.53	3.96	4.23
$M_s = 7$	28	3.35	4.57	3.39	**4.22**
$M_s = 8$	36	3.37	4.63	3.26	4.25
$M_s = 9$	45	3.39	4.68	3.24	4.33
number of substate rows relative to average duration					
$\alpha = 1.2$	12.0	3.34	4.55	7.49	4.65
$\alpha = 1.4$	16.2	3.31	4.52	4.68	4.27
$\alpha = 1.6$	21.3	3.30	4.52	3.72	**4.12**
$\alpha = 1.8$	26.9	3.35	4.59	3.41	4.16
$\alpha = 2.0$	33.4	3.37	4.66	3.28	4.20

Table 1. Word error rates [%] with full superstate connectivity

Table 2 shows the measured recognition performance following the different strategies of substate merging as discussed in 3.1. What we see is that it hardly has an impact on performance while it effectively reduces model sizes. The table lists the average number of substates per superstate and of connections between superstates.

In Table 3, we see word error rates measured with different degrees of superstate connectivity as discussed in 3.2. There is a obvious improvement through constraining the possible superstate transitions. When only allowing connections of rows of similar length we observe a negative impact from a too sparse connectivity. In case of normalizing the durations by the superstates' average durations and only allowing transitions between rows of similar normalized length, the very sparse connectivity achieves best

configuration	avrg. # of substates	avrg. # of connects.	trng. WER	test WER
$\alpha = 1.6$, as in Tab.1	21.3	33.5	3.72	4.12
acc. to left Fig.4	15.2	29.3	3.44	4.21
acc. to right Fig.4	15.2	19.6	3.43	4.20
$\alpha = 1.8$, as in Tab.1	26.9	43.0	3.41	4.16
acc. to left Fig.4	18.4	37.3	3.22	4.18
acc. to right Fig.4	18.4	23.0	3.23	**4.18**

Table 2. The impact of substate merging

test data performance. Interesting to see is that even the training data performance gains from strongly limiting the valid transitions.

avrg. # of connections	trng. WER	test WER	avrg. # of connections	trng. WER	test WER
connection of substate rows of similar length					
$M_s = 6$			$M_s = 7$		
36 (full)	3.96	4.23	49 (full)	3.39	4.22
19	3.75	4.10	24	3.23	4.12
16	3.70	**4.05**	19	3.23	**4.05**
11	3.40	4.47	13	3.23	4.43
connection of substate rows of similar normalized length					
$\alpha = 1.6$			$\alpha = 1.8$		
33.5 (full)	3.72	4.12	43.0 (full)	3.41	4.16
17.1	3.52	4.09	20.5	3.28	4.19
14.0	3.48	4.07	16.5	3.24	4.04
10.4	3.41	**4.04**	11.9	3.25	**4.02**

Table 3. Different degrees of superstate connectivity

With the configuration of narrow connectivity based on normalized durations, we have a 16% relative WER reduction over the baseline (without CDESHMM) as depicted in Table 1. It should be noted that this improvement is purely obtained through the modifications of the duration model.

Finally, Table 4 lists WERs achieved starting from the best set up of Table 3 and moderately relaxing the substate tying.

avrg. # of clusters	trng. WER	test WER	avrg. # of clusters	trng. WER	test WER
method 1a			method 2a		
3.3	3.06	3.68	3.2	3.20	3.71
4.0	3.04	**3.54**	3.9	3.19	3.63
4.7	3.08	3.62	4.8	3.35	3.69
method 1b			method 2b		
3.3	3.08	3.68	3.1	3.41	3.86
5.4	3.03	3.60	5.4	3.17	3.79
7.9	3.01	3.62	6.5	3.13	3.88

Table 4. Different methods and degrees of substate tying based on the $\alpha = 1.8$ and sparse connect. set up of Table 3 (WER 4.02%)

All proposed methods do yield some additional error reduction with the purely horizontal tying (1a) being the most effective configuration. In its best set up it contributes another 12% relative WER improvement. It is probably due to the large number of 10 (super-)states per digit HMM, that the feature stream is rather stationary within a state and that there is only minor improvement in the vertical clustering schemes.

Overall, the CDESHMM framework achieves a WER reduction from 4.79% to 3.54% which is a relative reduction of 26%.

Run-time and memory consumption

Computationally, CDESHMM are more demanding than ordinary HMMs. However, run-time benefits significantly from sparse superstate connectivity. Furthermore, the symmetric, strongly tied and loop-free internal superstate structure potentiates very efficient dedicated implementations. For now, all reported experiments were conducted using an untuned single-pass beam-search decoder with a fixed beam-width that only causes little pruning error. In this configuration, time and memory consumption increases with the proposed framework, but this increase is not as severe as one might expect. Decoding in the best set up of Table 4 (3.54% WER), for example, takes about 1.7 times as long as in the baseline set up without CDESHMM.

Duration probability scaling

In large vocabulary continuous speech recognition (LVCSR), it is common practice to weight language model scores against those derived from the acoustic model by a (well-tuned) language model scaling factor to compensate for different model quality and range. A duration model scaling factor can be introduced with a similar motivation [7]. Having a better duration model should allow a higher weighting. In this study though, no duration model scaling factor has been applied, i.e. it has always been 1.

6. CONCLUSION

The paper proposed and evaluated a highly constrained duration model as well as its integration into a standard first-order HMM-based speech recognizer. In the proposed model, the time-warping ability of HMMs which is usually realized through loopable states is replaced by a context-duration dependent bigram model which effectively limits the durational characteristics of valid paths through the HMMs. The model achieves best performance in a configuration free of self-loops and with thinned out superstate connectivity. In experiments on digit recognition it yields a 16% relative word error reduction. Combined with duration-dependent probabilistic distribution functions the relative improvement amounts to 26%. Future work will focus on higher order durational models than just bigrams, on augmenting the probabilistic distribution functions with the duration feature as well as on making effective use of the proposed framework in LVCSR.

7. REFERENCES

[1] V. R. R. Gadde, "Modeling Word Duration for Better Speech Recognition", Speech Transcript. Workshop, Maryland, 2000.

[2] F. Jelinek, "Statistical Methods for Speech Recognition", MIT Press, Cambridge, MA, Third Printing, 2001.

[3] M. Jones, P. C. Woodland, "Using Relative Duration in Large Vocabulary Speech Recognition", Eurospeech, Berlin, 1993.

[4] D. B. Paul, "New Results with the Lincoln tied mixture HMM CSR-System", DARPA Speech and Natural Language Workshop, Pacific Grove, 1991.

[5] D. Povey, "Phone Duration Modeling for LVCSR", ICASSP, Montreal, 2004.

[6] K. Power, "Durational Modelling for Improved Connected Digit Recognition", ICSLP, Philadelphia, 1996.

[7] J. Pylkkönen, "Phone Duration Modeling Techniques in Continuous Speech Recognition", Master thesis, Helsinki, 2004.

[8] M. J. Russell, R. K. Moore, "Explicit Modelling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", ICASSP, Tampa, 1985.

[9] D. Willett, C. Neukirchen, G. Rigoll, "A New Approach to Generalized Mixture Tying for Continuous HMM-Based Speech Recognition", Eurospeech, Rhodes, 1999.