Duration Modeling Techniques for Continuous Speech Recognition

Janne Pylkkönen and Mikko Kurimo

Neural Networks Research Centre
Helsinki University of Technology, Finland
janne.pylkkonen@hut.fi, mikko.kurimo@hut.fi

Abstract

Phone durations play a significant part in the comprehension of speech. The duration information is still mostly disregarded in automatic speech recognizers due to the use of hidden Markov models (HMMs) which are deficient in modeling phone durations properly. Previous results have shown that using different approaches for explicit duration modeling have improved the isolated word recognition in English. However, a unified comparison between the methods has not been reported.

In this paper three techniques for explicit duration modeling are compared and evaluated in a large vocabulary continuous speech recognition task. The target language was Finnish, in which phone durations are especially important for proper understanding. The results show that the choice of the duration modeling technique depends on the speed requirements of the recognizer. The best technique required a slightly longer running time than without an explicit duration model, but achieved an 8% relative improvement to the letter error rate.

1. Introduction

The modern automatic speech recognition (ASR) systems are based on modeling the phones with hidden Markov models (HMM), using HMM states in a left-to-right topology for each phone. The transition probabilities of the HMMs represent the statistical duration information of the phones. It has been noted that these transition probabilities have little effect to the recognition performance [1], and hence it is customary to ignore the use of more detailed durational information and rely more on the actual acoustic data.

The durational information is still worth of further examination. Although phone durations do not have actual discriminative role in English, they do help in distinguishing several words from each other, such as *sit* and *seat* or *ship* and *sheep*. In some other languages, for example in Finnish, phone durations can be the only clue in discriminating between certain words. Good duration modeling can therefore be a major issue.

It has been reported in several papers that using explicit state duration models with hidden Markov models improve the recognition accuracy [2, 3, 4]. However, most of the evaluations in these papers have been isolated word recognition tests with connected word models, not continuous speech recognition tests with phoneme based models nowadays in use. Besides, no single method have been found which would completely satisfy the modeling needs, and the different approaches have varying implications, for example, to the recognition efficiency. To gain more insight into this matter, this paper presents a comparison between three different extensions to integrate explicit duration models into the HMMs. The modeling techniques are evaluated using a modern phoneme based ASR [5] in a large vocabulary continuous speech recognition (LVCSR) task.

2. HMM based duration modeling techniques

Incorporating explicit state duration models into the HMMs introduces problems, as it breaks up some of the assumptions which are employed in the efficient HMM algorithms. A direct consequence of the Markov assumption is that state durations have a geometric distribution, defined by the probability of the self-transition. When this distribution is replaced with an explicitly defined one, the Markov assumption no longer holds. The Baum-Welch and Viterbi algorithms [6] used to find the optimal paths through an HMM heavily depend on this assumption, so they are no longer applicable in their basic forms. Modifying them to properly deal with the loss of this simplifying assumption seriously degrades their efficiency. The solution is then to find some other restrictive assumptions or to use sub-optimal algorithms.

Before reviewing the different duration modeling techniques, the distributions of phone durations are first examined.

2.1. Phone durations distribution models

For a phone model with three HMM states, the prior distribution of a phone duration is the convolution of three geometric distributions determined by the transition probabilities of the HMM. The properties of this prior distribution can be analyzed by considering the state durations as independent random variables. The mean and

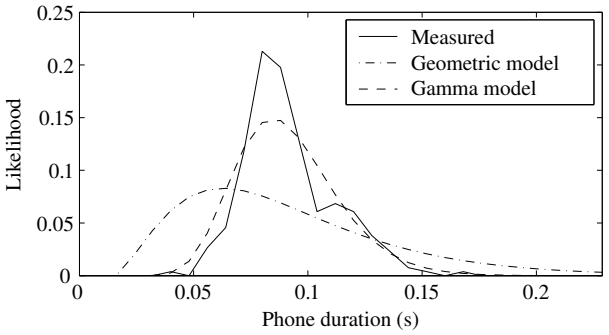

Figure 1: *An example of a phone duration distribution and models with convoluted state durations.*

variance of the overall distribution are therefore the sums of the means and variances of these random variables, respectively. This holds because the state durations really are independent from each other, due to the definition of HMMs. As a geometric distribution is defined by a single parameter, it defines both the mean and the variance of the distribution so that as the mean increases, so does the variance. The mean and the variance of the overall prior distribution are therefore closely coupled and restrict the form of the phone duration distribution.

The gamma distribution has been suggested as a good replacement for modeling state durations [7, 4]. It is a two-parameter distribution with an appealing shape for modeling duration information. Using gamma distributions the overall phone duration distribution is again a convolution of three distributions, but now with more freedom to adjust its shape. As an example, Figure 1 shows measurements of the durations of one triphone and two prior distributions for those durations, obtained from training the models. The other distribution utilizes the normal three-state HMM with geometric state durations, while the other has gamma distributions fitted to the state durations. Even though convoluting the three geometric state duration distributions permits a gamma-like overall distribution, it fits clearly worse than the convolution of the three gamma distributions.

For phone durations to be useful as an information source in ASR, it would be desirable that they contained only moderate variation. Unfortunately several factors affect the duration of phones, such as stress, the location of word and syllable boundaries, the number of syllables in a word, the phoneme context, and the overall speaking rate [8]. In this work, only the effect of phoneme context have been taken into account by modeling the durations of different triphones separately. Some examples of adapting duration models to speaking rate can be found from [9].

2.2. Hidden semi-Markov models

If a normal HMM is extended by explicitly defining the state duration distributions, the resulting model is called a *hidden semi-Markov model* (HSMM) [4]. In such a definition the self-transition probabilities are ignored and the state occupancy is defined by a state duration distribution. As mentioned above, this kind of definition violates the Markov assumption, as the transition probabilities at any time depend on the time the process has remained in the present state. When considering, for example, the Viterbi algorithm, this implies that it is no longer enough to store the state probabilities for one time step, but a complete state probability history is needed.

The easiest way to relieve the computational burden is to define maximum state duration D. This way the state probability history is needed only for D time steps, and the algorithm suffers only a slow down by factor D. However, a reasonable value for D is on the order of 25 frames [6], which already results in a serious degradation in efficiency. Bonafonte *et al.* [4] presented a pruning theorem with which the search space of the Viterbi algorithm can be further limited without compromising the optimality of the algorithm. They reported an increase of computational effort of about 3.2 times with respect to conventional HMM, the increase being almost independent of the actual value of D. The only assumption their pruning theorem requires is that the state duration distributions must be log-convex, as is the case for most parametric distributions useful for the purpose [4]. In particular, the gamma distribution can be used if it is restricted to have a mean greater than its standard deviation. This was found to be fulfilled in all practical cases [4], so it should not constrain the use of gamma distribution in duration modeling. For the evaluation in this work, the HSMMs were implemented in the ASR system using the above-mentioned pruning theorem and gamma distributed state durations.

2.3. Expanded state HMM

Markov models can be made to approximate general distribution functions. As the acoustic models already rely on Markov models, it is possible to include more flexible duration distributions directly to the HMM framework. This can be achieved by expanding each HMM state to a sub-HMM, which shares the same emission probability density and realizes the correct state duration distribution with its topology and transition probabilities. This kind of model is called the *expanded state HMM* (ESHMM) [3].

When constructing such a model it is important to note that the Viterbi algorithm used in recognition does not simulate the Markov model in a strictly mathematical way. That is, it does not sum over all the possible paths, but finds only one path over which it computes the probability. This restricts the usable topologies for the sub-HMMs [3]. Figure 2 shows a topology suggested in [10]. By introducing a self-transition to the end of the sub-HMM, there is no need to explicitly restrict the max-

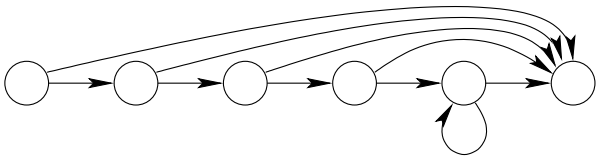

Figure 2: *Sub-HMM topology. The rightmost state illustrates the next HMM or sub-HMM state, so that the minimum duration in the sub-HMM is one.*

imum duration of one normal HMM state.

Expanding each HMM state to this kind of sub-HMM introduces a large number of free parameters to be estimated. It may be therefore necessary to constrain the parameters in some way. In [10] the number of states in sub-HMMs were determined by the number of occurrences in the HMM state in the training phase, and the transition probabilities of all the sub-HMM states of all the phone models were set to be the same. In this work, a heuristic rule for determining the number of sub-HMM states was used so that good fits to the measured duration distributions were achieved with low numbers of sub-HMM states. On average, the sub-HMMs had 3.8 states. The transition probabilities were constrained by fitting a gamma distribution to the measured duration distribution.

2.4. Post-processor duration model

Both HSMM and ESHMM degrade the efficiency of the recognition, the former by altering the algorithms and the latter by introducing additional states for the HMMs. Juang *et al.* [2] proposed a duration model which avoids this kind of loss of efficiency. Their method uses the output of the Viterbi algorithm and ranks the proposed paths using better models for the state durations. The method is therefore called the *post-processor duration model*. The augmentation of the log likelihood given by the Viterbi algorithm can be stated as

$$\log \hat{f} = \log f + \alpha \sum_{j=1}^{N} \log d_j(\tau_j). \qquad (1)$$

f denotes the likelihood score given by the Viterbi search, α is an empirical scaling factor, N is the number of distinct HMM states through which the best path traversed, d_j are the duration probability distribution functions of those states, and τ_j are the durations spent in each state.

3. Evaluation

3.1. Setup

The utility of the duration modeling techniques was evaluated with speaker dependent speech recognition tests. Finnish was used as the target language, as the proper understanding of Finnish is more dependent on the correct durational information than, for example, with English. The speech material was a book spoken by a professional speaker, which was a reasonable choice for minimizing the unwanted variation of phone durations. An extract of 12 hours was used to train the models, and independent parts of 9 and 30 minutes were used as development and evaluation sets, respectively. The development set was used to optimize the empirical scaling factors for the log likelihoods of the language model, the transition probabilities, and the duration distribution probabilities.

The speech recognition system used for the evaluation has been presented in [5]. The number of triphone models was empirically adjusted to the available data. For the language model, a morph based trigram model was used. All the duration distributions were modeled with gamma distributions. As the different duration modeling techniques affect both the efficiency and the accuracy of the recognition, the recognition tests were run with different pruning settings (affecting the optimality of the Viterbi algorithm) to achieve different running times. The running time is indicated by a real-time factor, which should be interpreted only as a relative value for the number of reasons affecting the actual speed of the recognition. The recognition accuracy was measured by a letter error rate (LER). As compared with the word error rate (WER), it is more suitable for a language such as Finnish where rather long words consisting of many morphemes are common.

3.2. Results

Figure 3 shows the recognition accuracy as a function of the real-time factor for different setups. The model labeled as "HMM" is the baseline result without explicit duration modeling. The figure shows clearly that the intended running speed affects the choice of the best duration modeling technique. For moderate speeds (real-time factors 10 to 30) the post-processor model functions best. But if the pruning level of the recognition is set to low enough, the HSMM outperforms the others. The ESHMM does not seem to produce good results, despite its intuitive approach.

All the models suffer from random fluctuation in the LER measurements after they have reached their optimal running speed. This is due to inherent noise in the measurements, along with the effects resulting from the general pruning strategy used in the Viterbi algorithm. Measured from the points of the best performance, the letter error rate of the post-processor duration model was 2.73%, corresponding to the word error rate of 15.3%. The HSMM achieved a LER of 2.63% (WER 15.2%). Compared to the baseline result with a LER of 2.88% (WER 16.2%), the post-processor duration model improved the LER about 5%, while the HSMM achieved about 8% relative improvement.

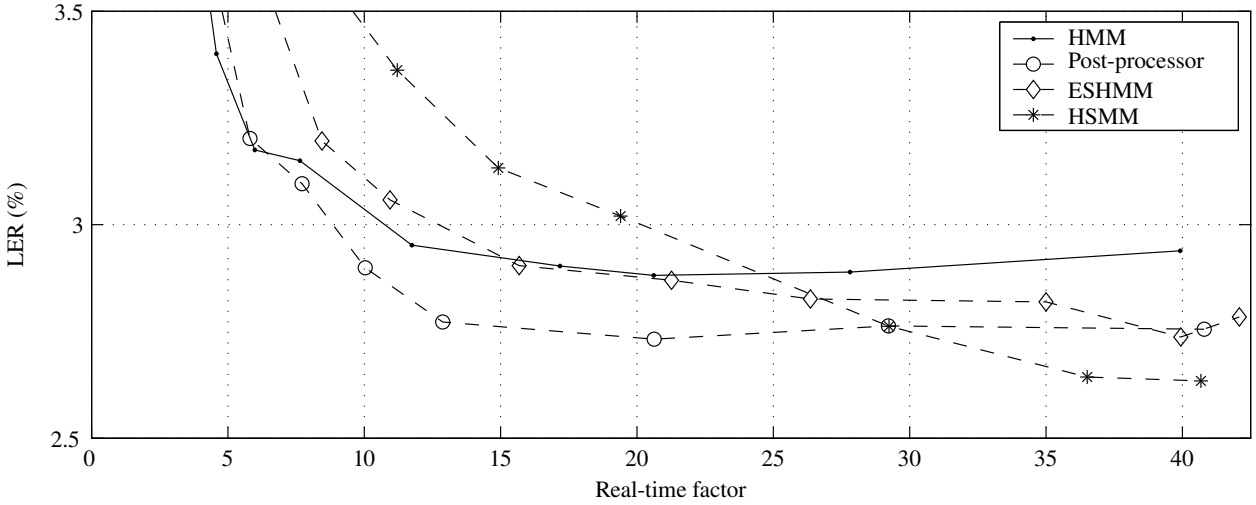

Figure 3: *Comparison of the performance of different duration modeling techniques.*

4. Conclusions

This paper presented a comparison between three different techniques for improving the phone duration models in an LVCSR task. Depending on the efficiency requirements, either simple post-processor duration model or a more complex hidden semi-Markov model based approach was shown to give the best results. The former is easy to be implemented and works well with moderate running speeds. The latter requires modifying the Viterbi algorithm, and it slows down the recognition. However, it achieved the best recognition accuracy with a statistical significant 8% relative improvement to the letter error rate when compared to the normal HMM based system without explicit duration modeling.

5. Acknowledgements

This work was supported by the Academy of Finland in the projects *New information processing principles* and *New adaptive and learning methods in speech recognition*. We thank the Finnish Federation of the Visually Impaired and the Departments of Phonetics and General Linguistics of the University of Helsinki for providing the speech data. We also thank the Finnish news agency (STT) and the Finnish IT center for science (CSC) for the text data.

6. References

[1] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, no. 3, pp. 205–231, May 1996.

[2] B. H. Juang, L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition," in *Proc. ICASSP*, 1985, pp. 9–12.

[3] M. J. Russell and A. E. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition," in *Proc. ICASSP*, 1987, pp. 2376–2379.

[4] A. Bonafonte, X. Ros, and J. B. Mariño, "An efficient algorithm to find the best state sequence in HSMM," in *Proc. Eurospeech*, 1993, pp. 1547–1550.

[5] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proc. Eurospeech*, 2003, pp. 2293–2296.

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, February 1989.

[7] S. E. Levinson, "Continuously variable duration hidden Markov models for speech analysis," in *Proc. ICASSP*, 1986, pp. 1241–1244.

[8] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[9] K. Samudravijaya, S. K. Singh, and P. V. S. Rao, "Pre-recognition measures of speaking rate," *Speech Communication*, vol. 24, pp. 73–84, 1998.

[10] A. Noll and H. Ney, "Training of phoneme models in a sentence recognition system," in *Proc. ICASSP*, 1987, pp. 1277–1280.