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


