Using Phone Durations in Finnish Large Vocabulary Continuous Speech Recognition

Janne Pylkkönen and Mikko Kurimo

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT
FINLAND
E-mail: janne.pylkkonen@hut.fi, mikko.kurimo@hut.fi

ABSTRACT

Finnish is one of the languages where phone durations discriminate between words. They therefore have significant meaning in the proper recognition of speech. Modern large vocabulary continuous speech recognizers do not offer reasonable means to model these durations, which would be necessary in order to seamlessly deal with such a language. Therefore some explicit actions have to be taken to be able to distinguish certain words from each other as the only cues for doing this might be prosodic ones, namely the durations. In this work, an extension of an existing speech recognition system to include models for discriminatively important phone durations is studied. The explicit duration model applied resulted in 5% relative reduction in the letter error rate of the recognition task.

1. INTRODUCTION

Finnish is an example of a language in which phone durations have discriminative role in the speech. The single and double phonemes are distinguished from each other by different durations of the corresponding phones. The same is apparent also in the written forms of the words, making it easy to analyze the phenomenon. For example, words *tuli* and *tuuli* (fire and wind) have the same phones in the spectral sense, but the vowel /u/ has increased duration in the latter word. This kind of distinction between words is not an exception, many similar word pairs exists in Finnish.

The discriminative meaning of phone durations in Finnish is in contrast to their use in some other languages, for example, in English. The other extreme would be that the durations of the phones had no role whatsoever with the identification of the words. But durations and acoustic quality can also be coupled, as is the case in English. An example of a word pair in which this occurs is *seat* and *sit*. The middle part of the latter word is clearly shorter than that of the former, but there is also difference in the spectral contents of the phones. However, in English, the spectral cue is the more important for discrimination than the duration [1].

From this consideration it is apparent that for some languages, including Finnish, the phone duration modeling is crucial for the proper recognition of speech. But even with languages where phone durations do not give discriminative information between the words, analyzing them can provide useful information to aid the recognition task. Unfortunately, the modern speech recognition systems generally model the phone durations rather poorly due to use of hidden Markov models (HMMs) as their underlying acoustic model. Several alleviations have been proposed [2, 3, 4], but the field seems to lack an established method to deal with the problem.

In this work, the use of explicit phone duration models is studied in the context of an existing speech recognition system. To alleviate the limitations of HMMs, a duration model functioning as a post-processor to the path comparison in the normal decoder was implemented. This kind of duration model is easy to be implemented and efficient to be used, yet powerful enough to improve the accuracy of the recognizer.

2. PHONE DURATION MODELING

2.1 HMM Based Phone Duration Models

Hidden Markov models have intrinsic geometric state duration distributions, as the duration of one HMM state is completely determined by the probability of a self transition [5]. In majority of the modern phoneme based speech recognizers a phone is modeled as a three-state left-to-right HMM. The resulting phone duration distribution is therefore a convolution of these three geometric distributions, when considered prior to the acoustic information. Although this distribution has three free parameters, they are coupled in such a way that the overall distribution is insufficient to model the phone durations properly [6].

Figure 1 shows an example of the poor modeling of the phone durations with the normal three-state HMMs. The solid line shows an unsmoothed duration distribution of a triphone, where /s/ has an /a/ as the left context and /t/ as the right context. The data was measured over 879 occurrences of the triphone in a speech material spoken by the same speaker. The dash-dotted line shows the convolution of the three geometric state duration distributions, therefore representing the phone duration distribution. The fit is far from the objective distribution. On the other hand, the dashed line shows the convolution of three gamma distributions fitted to model the same HMM state durations.

The fit is now almost perfect. This suggests that gamma distributions are much more suitable for modeling the state durations, and they have indeed been reported to be used for that purpose [3]. They are also the choice for the duration distributions in this work.

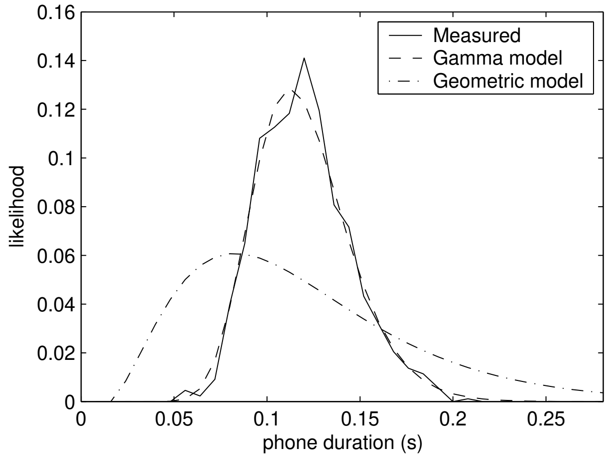

Fig. 1. An example of a phone duration distribution and models with convoluted state durations, as explained in the text.

Several methods have been proposed to enable the use of more general state duration distributions with HMM based models. Extending the HMM directly to include an explicit duration model instead of an intrinsic one leads to so called *hidden semi-Markov models* [7, 3, 8]. Another method to better model the state durations employs the fact that Markov models can be used to model general probability distributions [9]. Therefore HMM states can be extended to sub-HMMs, sharing the same acoustic emission density, to explicitly model the state durations. Resulting model is called the *expanded state HMM* [4]. Unfortunately, both of these models degrade the recognition efficiency, thus cutting the benefits of enhanced duration modeling.

Juang *et al.* [2] proposed a duration model which avoids the loss of efficiency in the recognition. Their method uses the output of the Viterbi search [5] at the heart of the decoding (recognition) process, and ranks the proposed paths using better models for the state durations. The method is therefore called the *post-processor duration model*. The original version of this method was formulated as an augmentation to the Viterbi search likelihood:

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

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.

The post-processor method can also be derived from the following probabilistic considerations. If simplified, the decoder can be seen to pick the phoneme sequence W for which the likelihood $p(\boldsymbol{O} \mid W, \lambda_a, \lambda_d)$ of the acoustic frame sequence $\boldsymbol{O} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_t$ is the highest. The decoder is given the acoustic and duration models as a prior knowledge, denoted by λ_a and λ_d, respectively. The acoustic likelihood can be computed as a sum over the paths of the HMM forming the phoneme sequence W:

$$p(\boldsymbol{O} \mid W, \lambda_a, \lambda_d) = \sum_{Q_i \in Q} p(\boldsymbol{O} \mid Q_i, \lambda_a) P(Q_i \mid W, \lambda_d), \tag{2}$$

where Q is the set of valid HMM state sequences. In the Viterbi search used in decoding, the likelihood is approximated with a single best path:

$$p(\boldsymbol{O} \mid W, \lambda_a, \lambda_d) \approx p(\boldsymbol{O} \mid Q_{best}, \lambda_a) P(Q_{best} \mid W, \lambda_d). \tag{3}$$

The first term of the right hand side is simply the acoustic probability given the state sequence Q_{best}. In the context of HMMs, the second term is the product of transition probabilities for the state sequence Q_{best}. However, this term could be computed more correctly, if we forget the HMM context and adopt more accurate state duration models. Therefore the resulting likelihood can be computed as a product of the acoustic likelihood and the duration likelihood, which both are computed over the best path given by the Viterbi search. When dealing with log likelihoods, this results in a sum of logarithms of the acoustic and duration likelihoods. This is now in accordance with the equation 1, except that the f in Eq. 1 includes, along the acoustic likelihood, also the transition probabilities of the HMM path, which is the duration information with the geometric state distributions. The implications of this depend on the scaling of the acoustic likelihoods and on the scaling factor α.

Tests with different duration modeling methods showed that the post-processor method gave the best results in relation to the recognizer efficiency [6]. Although it is not mathematically as justified as the other methods, it benefits for having almost negligible impact to the recognition efficiency. The decoder can therefore be run with more time consuming parameters to achieve the performance gain. The performance and implementation of the post-processor method is deeply coupled with the structure of the decoder, which is presented in the next section.

2.2 The Stack Decoder Based Speech Recognition System

The speech recognition system at HUT utilized for this study has been presented in [10]. The specialty in the system is the use of morphs instead of words as the language model units. The morphs were chosen to model Finnish, which has huge number of word forms due to extensive use of suffixes and compound words. These morphs are morpheme-like units which are discovered in an unsupervised manner, thus achieving language independence and avoids the need of coding extensive morphological rules.

The decoder of the speech recognizer [11] is based on the principle of stack decoding [12]. The existing recognition hypotheses are expanded with new words (or in this case morphs) starting from each time instance where some previous hypothesis ends. Several hypotheses are stored in stacks and expanded appropriately with new word (morph) alternatives. These new alternatives are searched with a local Viterbi search in a window of about 1-2 seconds. The breadth of the Viterbi search can be controlled with a so called beam parameter, which defines how much the log likelihood of different expansion alternatives are allowed to deviate from the best one. The effect of this pruning is in close correlation to the speed of the decoder, so it can be used as a mean to control the tradeoff between the accuracy and efficiency. The larger the beam parameter is, the more accurate recognitions can be achieved.

The search strategy which the decoder employs suits well to the post-processor duration model. As several competing hypothesis paths are kept as a starting point for new hypotheses, many path alternatives are presented to the post-processor model. This makes it more probable that the correct paths are evaluated using the better duration models. Ranking the new hypotheses with better models leads to more correct decisions and guides the recognition to more accurate results. Furthermore, due to the expansion scheme used in the decoder several expansions may result the same hypothesis, but with slightly differing paths over the HMM state lattice. This also reduces the possibility that the best path in respect to the better duration models would be missed.

The decoder combines the different knowledge sources by summing their log likelihoods together with appropriate scalings to get the final likelihood estimate for each hypothesis. These knowledge sources are the acoustic likelihood (HMM emission probabilities), the HMM transition probabilities, the new duration model (the latter term of the Eq. 1) and the language model. The log likelihoods of the three last sources are scaled with respect to the acoustic log likelihood to achieve the best recognition performance. During the experiments, all these three parameters were optimized with a development set independent of the actual test set. The actual values of the scaling factors depend on the scaling of the acoustic log likelihood defined by the Gaussian mixture models. In this system these acoustic probabilities were normalized to sum to one for each acoustic frame.

3. EXPERIMENTS

The models used for the speech recognition experiments were speaker dependent triphone models trained from an about 12 hour extract of a Finnish book spoken by one female reader. Independent parts of 9 and 30 minutes of the same material were used as development and test sets, respectively. The number of different triphones was empirically adjusted to the available data. Each triphone was modeled with three-state left-to-right HMM model, each HMM state having Gaussian mixture emission density with four Gaussian components. For the explicit state duration models, gamma distributions were used.

For the recognition tests, letter error rates (LER) were used as a criterion for the recognition performance. There are number of reasons for this decision. Word error rate (WER), which is common in speech recognition measurements, is not well applicable for Finnish where rather long words consisting of many morphemes are common. Word error rate also penalizes too much for misrecognized word breaks. This becomes an issue when language model units are not words. Letter error rate, on the other hand, is a good and meaningful criterion if the recognition result is intended to be manually corrected or read by humans. The actual task for which the speech recognizer is designed to naturally determines the proper error measure, and this latter argument has been seen as the most relevant one in the current system.

The recognition evaluation was performed with different decoder beam values, resulting in different running times. This way a recognition performance as a function of recognition speed could be obtained as in figure 2. The speed is indicated by the real-time factor of the decoding, and it should be interpreted only as a relative value, for the number of reasons affecting the actual speed of the recognition.

Four different models were evaluated. The baseline did not distinguish between single and double phonemes. These were separated to own acoustic models to improve the recognition accuracy. Then upon that two versions of post-processor duration models were implemented and tested. The first (1) does not include the transition probabilities encountered during the Viterbi search, whereas the second (2) does, thus implementing exactly the equation 1. Table 1 shows the lowest letter error rates of the evaluation, along with the corresponding word error rates for reference.

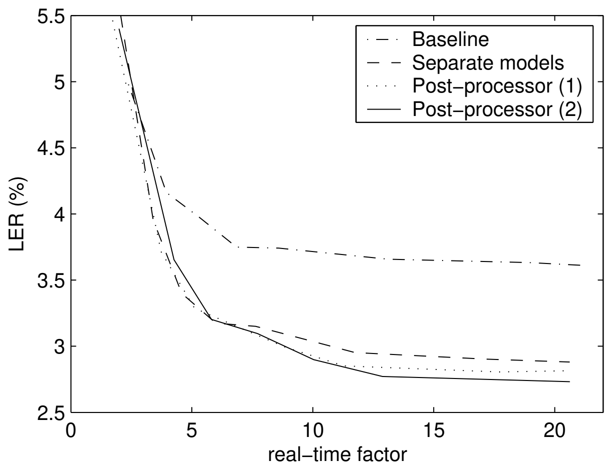

Fig. 2. Speech recognition results with various running times

Table 1. Speech recognition results for real-time factor 21

Model	LER (%)	WER (%)
Baseline	3.61	19.8
Separate phonemes	2.88	16.2
Post-processor (1)	2.81	15.6
Post-processor (2)	2.73	15.3

4. DISCUSSION

The system as reported in [10] did not distinguish between different lengths of phonemes and it was entirely up to the language model to decide which variant of the phoneme was more correct in the context of a hypothesis. Thus a simple separation of the single and double phonemes alone improved the recognition accuracy significantly. This way the different durations of phones can be modeled using the HMM transition probabilities. Although there should not be significant difference in the spectral contents of the phones corresponding to the single and double forms of the same phoneme, the acoustic emission densities were kept separate to keep the system as simple as possible.

It can be noted from Fig. 2 that the better modeling of the phone durations begin to have influence only after a certain accuracy has been achieved, in respect to the pruning of the decoder. Also what is seen in the figure is that the post-processor method performs slightly better when implemented as originally proposed, by augmenting the log likelihood given by the Viterbi search. This suggests that the transition probabilities may carry some additional information not available in the gamma distributed state duration models. With this duration model a letter error rate of about 5% smaller (6% smaller WER) is achieved than with the same acoustic models but without the post-processor. The model without separate models for single and double phonemes is clearly the worst.

The assumptions with the implemented post-processor model are that we get the correct paths from the Viterbi search and that by ranking them using better duration models really gives us more information to aid in choosing the best hypothesis. The evident problem results from the former assumption, as the best path relative to the better duration models needs not be the same as the one Viterbi search finds using the simple geometric duration models. However, the evaluations show the the post-processor model works well despite this inconsistency.

5. CONCLUSIONS

Explicit modeling of phonemes of different lengths, and therefore the durations of phones, is crucial in Finnish to distinguish certain words from each other. Better models can be built if the durations of the phones are modeled more accurately than the standard HMM framework allows. During the tests with different phone duration models it was noted that a simple post-processor duration model was the best with respect to the recognition efficiency. The evaluation reported in this work indicate that about 5% relative reduction in letter error rate can be achieved with this model. It should be noted, however, that the actual performance of the model depends on the decoder in which the model is implemented.

REFERENCES

[1] K. Wiik, *Fonetiikan perusteet*, 2nd ed. WSOY, 1998, (in finnish).

[2] B. H. Juang, L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "Recent developements in the application of hidden Markov models to speaker-independent isolated word recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 9–12.

[3] S. E. Levinson, "Continuously variable duration hidden Markov models for speech analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1986, pp. 1241–1244.

[4] M. J. Russell and A. E. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1987, pp. 2376–2379.

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

[6] J. Pylkkönen, "Phone duration modeling techniques in continous speech recognition," Master's thesis, Helsinki University of Technology, 2004.

[7] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 5–8.

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

[9] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *Journal of Acoustic Society of America*, vol. 83, no. 4, April 1988.

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

[11] T. Hirsimäki, "A decoder for large vocabulary continuous speech recognition," Master's thesis, Helsinki University of Technology, 2002.

[12] D. Willett, C. Neukirchen, and G. Rigoll, "DUcoder-the Duisburg University LVCSR stackdecoder," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000, pp. 1555–1558.