Word Duration Modeling for Word Graph Rescoring in LVCSR

Dino Seppi[1]*, Daniele Falavigna*[1]*, Georg Stemmer*[2]*, Roberto Gretter*[1]

[1] Fondazione Bruno Kessler - irst, *formerly ITC-irst*, Trento, Italy
[2] Siemens AG, Corporate Technology, Munich, Germany

Abstract

A well-known unfavorable property of HMMs in speech recognition is their inappropriate representation of phone and word durations. This paper describes an approach to resolve this limitation by integrating explicit word duration models into an HMM-based speech recognizer. Word durations are represented by log-normal densities using a back-off strategy that approximates durations of words that have been observed seldom by a combination of the statistics of suitable sub-word units. Furthermore, two different normalization procedures are compared which reduce the influence of the implicit HMM duration distribution resulting from the state-to-state transition probabilities. Experiments on European parliamentary speeches in English and Spanish language show that the proposed approaches are effective and lead to small, but consistent reductions in the word error rate for large-vocabulary speech recognition tasks.
Index Terms: hidden Markov models, duration modeling, speech recognition, stochastic approximation

1. Introduction

Hidden Markov Models (HMMs) constitute the most frequently used framework for an efficient representation of speech signals. HMMs allow to model the time-dependent variability of speech within a rigorous, computationally efficient mathematical paradigm. However, they also suffer from some limitations. One of these is the inappropriate modeling of phone and word durations: HMMs model state durations by geometric probability density functions (pdf).

Early attempts to improve duration modeling are *Hidden Semi-Markov Models* (HSMM), *Expanded State HMM* (ESHMM), and the *Post-processor duration model* (PPDM). HSMM replaces HMM self state transition probabilities with an explicit state duration pdf; this approach has been repeatedly evaluated using implementations proposed by Levinson [1] and Fergusson [2]. Both methods were reviewed and compared in [3] with regard to isolated word recognition tasks, and in [4] for Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. However, in spite of a considerable increase of the computational burden and the need of non-straightforward changes in the decoding procedure, the performance increase is moderate. Therefore common state-of-the-art Automatic Speech Recognizers (ASR) do not implement such facilities. The PPDM has been first proposed in [5], and later adapted to LVCSR tasks in [4]. This approach avoids the performance degradation by simply rescoring the paths of the trellis before the standard Viterbi search algorithm is applied. The rescoring is accomplished augmenting the original likelihood by weighted duration probabilities. This idea has been further improved in [6], where the probabilities of the N-best recognized strings were rescored applying more accurate word duration models, and more recently in [7], where N-best ASR output was replaced by Word Graphs (WG). This last approach, also adopted in our work, is relatively easy to implement since common LVCSRs are based on a multi-pass decoding strategy: the WGs generated in the last step are suitable for further elaboration. This allows to include in the decoder acoustic information spanning time intervals much longer than those spanned by HMM states. Therefore, the proposed approach could be adopted for exploiting other types of *supra-segmental* information beside duration, such as pitch contours, energy contours, etc.

In this paper we propose a simple word duration model that employs just a single word-based pdf, namely the *log-normal* pdf. Other approaches like [7, 6] evaluate word duration probabilities through a set of Gaussian Mixtures (GMs): one GM is used for each phone sequence of a given word. The main advantage of our method is that it allows to discard the phone segmentation during the decoding phase. At the same time word duration statistics can be estimated robustly using the back-off strategy described in the remainder of this paper, a strategy which is difficult to apply when using GMs as in [7].

When integrating an explicit word duration model with conventional HMMs it is important to take into account the influence of the intrinsic duration pdf associated to state-to-state transition probabilities of the HMMs. Therefore we introduce and compare two different word duration normalization methods that partially remove the contribution of the intrinsic duration pdf. To evaluate the benefits of our approach we performed some LVCSR experiments over the TC-STAR European Parliament Plenary Sessions (EPPS) tasks. These include both English and Spanish training and test speech data.

The paper is organized as follows: Section 2 describes the word duration model and the context where it is applied; Section 3 presents the ASR system and the decoding procedure based on word graph rescoring; in Section 4 experiments are described and preliminary results reported. Section 5 concludes the paper and gives a short outlook on our future work.

2. Duration modeling

In the following description we denote a word sequence $w_1, .., w_M$ with W, a sequence of acoustic feature vectors with A, and a word duration sequence $d_1^w, .., d_M^w$ with D. The standard approach for speech recognition searches for the word string W^* corresponding to the best-scored path inside a network of HMMs. For the combination of the acoustic and linguistic scores $P(A|W)$ and $P(W)$ the Bayes-rule is employed:

$$\begin{aligned} W^* &= \underset{W}{\operatorname{argmax}}\, P(W|A) = \underset{W}{\operatorname{argmax}}\, \frac{P(A|W)P(W)}{P(A)} \\ &= \underset{W}{\operatorname{argmax}}\, P(A|W)P(W) \end{aligned} \quad (1)$$

Usually the Viterbi algorithm is applied, which not only generates the best-scored word sequence W^* but also the corre-

sponding word durations D^*. Thus, we can introduce the duration sequence D as additional word-level information into Eq.1, which can be reformulated as:

$$\begin{aligned} W^*, D^* &= \underset{W,D}{\text{argmax}}\, P(W, D|A) \\ &= \underset{W,D}{\text{argmax}}\, \frac{P(A, D|W)P(W)}{P(A)} \\ &= \underset{W,D}{\text{argmax}}\, P(A, D|W)P(W) \quad (2) \\ &= \underset{W,D}{\text{argmax}}\, P(A|W, D)P_{HMM}(D|W)P(W) \quad (3) \end{aligned}$$

where the last line has two main differences compared to the conventional approach: firstly, there is an additional term $P_{HMM}(D|W)$ called the *HMM implicit duration model*. This corresponds to a factorization of the conventional acoustic density $P(A, D|W)$, i. e. the acoustic model generates a sequence of acoustic vectors with a certain duration, into $P(A|W, D)$ and $P_{HMM}(D|W)$. Secondly, $P(A|W, D)$ contains a dependence of the acoustic features A on the word durations. Assuming that, given a word sequence W, acoustic features A could be considered as independent from word durations D [8], we would obtain a duration independent acoustic model $P(A|W)$. Some studies have demonstrated that it is possible to gain in ASR performance by customizing HMMs according to D, e.g. by taking into account the speaking rate [9, 10]. However, in this study we will ignore the influence of the duration on the feature vectors. Thus the decoding strategy might be drawn by Eq. 4, where the dependence of $P_{HMM}(D|W)$ upon the HMMs chain characteristics is highlighted by the subscript *'HMM'*:

$$W^*, D^* \approx \underset{W,D}{\text{argmax}}\, P(A|W)P_{HMM}(D|W)P(W) \quad (4)$$

In the next subsections we describe how to eliminate the 3-state geometrical distribution $P_{HMM}(D|W)$ by substituting it with a more reliable log-normal distribution $P_{LN}(D|W)$. We show that this is not straightforward as $P(A, D|W)$ is not factorized and therefore $P_{HMM}(D|W)$ must be estimated separately.

2.1. Log-normal duration model

The characteristics of HMM mentioned above suggest that they do not supply a valid framework for correct modeling of word and phone durations. As a matter of fact, it is well known [11, 7] that duration statistics are better approximated by gamma, mixtures of gaussians (GM), and log-normal pdfs. Here we opt for the log-normal choice. Compared to the gamma pdf, the log-normal pdf presents two major advantages: its maximum likelihood estimation is available in closed form, while the gamma estimation needs numerical methods, and it has been proved to model phone and word duration statistics in a more reliable way [11]. At the same time, the log-normal pdf can be preferred to the GM pdf because the former requires a smaller number of parameters to be estimated, it is generally less prone to overfitting, and it allows easy back-off methods for modeling a word pdf through phone ones.

The estimation of the model parameters in the case of phone duration statistics is quite easy because of the relatively small number of phones in each language and the large number of different realizations available in speech databases. On the contrary, when dealing with word duration statistics, problems arise due to the arbitrarily huge number of words: it is impossible to define a vocabulary able to cover all possible utterances, if not because of the presence of proper names. The problem is thus to back-off from word duration statistics to phone duration statistics. This task is not straightforward because of many reasons: the same phone in different positions within a generic word may be uttered with different speeds due to sentence accent, lexical stress, phone context, word position in the sentence, e.g. word proximity to a pause, word length, etc. Inter-speaker variability is another important factor.

The adopted back-off approach has been borrowed from the wireless communications field, where there is the need to compute the total co-channel interference, usually modeled as a sum of log-normal random distributed signals. Likewise, consider the durations of N phones $d_1^p, .., d_N^p$ of word w. Given that random variable represented by the duration d^w of word w is equal to the sum of the random variables representing its phones, $d_1^p + .. + d_N^p$, and assuming that the sum is well approximated again by a log-normal distribution [12, 13], it is possible to find the first moments of d^w from the moments of $d_i^p, \forall i = 1..N$. Here the Wilkinson-Fenton's algorithm is used. It basically relies on the hypothesis mentioned above and proceeds by matching of the first and second moments, thus giving the best approximation to the exact sum in the middle range of the random variable [13]:

$$\mu(d^w) = 2\ln u_1 - 1/2 \ln u_2 \quad (5)$$

$$\sigma(d^w) = \sqrt{\ln u_2 - 2\ln u_1} \quad (6)$$

where

$$u_1 = \sum_{i=1}^{N} \exp\left(\mu(d_i^p) + \frac{\sigma^2(d_i^p)}{2}\right) \quad (7)$$

$$\begin{aligned} u_2 = &\sum_{i=1}^{N} \exp\left(2\mu(d_i^p) + 2\sigma^2(d_i^p)\right) \\ &+ 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \exp\left(\mu(d_i^p) + \mu(d_j^p)\right) \\ &\times \exp\left(\frac{1}{2}(\sigma^2(d_i^p) + \sigma^2(d_j^p) + 2r_{ij}\sigma(d_i^p)\sigma(d_j^p))\right) \quad (8) \end{aligned}$$

where $\mu()$ and $\sigma()$ are the two parameters that identify the log-normal pdf [1], and r_{ij} is the correlation coefficient of couples of adjacent phones of lengths d_i and d_j. This allows to model with diphthongs and phone relations in general, without the strong limitation of assuming that durations of phone pairs are independent. Once the back-off strategy, which takes place when the occurrences of word w are less than 20, has been fixed we could finally model each word by an unique pdf, namely $p_{LN}(d|w)$[2].

2.2. HMM duration model normalization

Replacement of implicit word duration is achieved by normalizing the acoustic model of Eq. 2 by an estimate of $P_{HMM}(D|W)$. As a matter of fact $P_{HMM}(D|W)$ is merged into the acoustic model $P(A, D|W)$ and its precise evaluation would require a throughout inspection of the best-scored HMM state sequence exploited by Viterbi decoding. Storing this path for each word hypothesis would partly compromise the recognizer efficiency. Therefore we approximate $P_{HMM}(d|w)$ for each word w, given a generic word length d, in the following two alternative ways.

[1] μ and σ^2 are first and second moment of the variable's *logarithm*.

[2] Henceforth probabilities are denoted by $P()$, while probability density functions by $p()$.

The first alternative, namely $P^{\mathsf{F}}_{HMM}(d|w)$, is computed as follows: for a left-to-right HMM chain containing I states $s = 1, .., I$ the following equation holds:

$$P^{\mathsf{F}}_{HMM}(d|w) = \sum_{S} \prod_{t=1}^{d-1} P(s_{t+1} = i | s_t = j) = \sum_{S} \prod_{t=1}^{d-1} a_{ji}$$

where S denotes w's state sequence, a_{ji} the transition probability from state j to state i, and s_t the state at time t. $P^{\mathsf{F}}_{HMM}(d|w)$ has been recursively evaluated using the *forward algorithm*. Note that the *Viterbi algorithm*, on the contrary, suffers from the heavy limitation of getting stuck for most of the time on the HMM self-transition probability a_{ii} with the highest value in the word. For this study, $P^{\mathsf{F}}_{HMM}(d|w)$ has been computed off-line using plausible word lengths and the transition matrices of the HMMs used in the second recognition pass (Section 3).

The second alternative relies on the approximation of each phone being visited for a time proportional to its self-transition probabilities a_{ii}. $P^{\mathsf{P}}_{HMM}(d|w)$ is directly obtained from the transition matrices using the following equation:

$$P^{\mathsf{P}}_{HMM}(d|w) = \prod_{i=1}^{I} a_{ii}^{\delta_i - 1} \cdot (1 - a_{ii}) \quad (9)$$

where δ_i is the duration in the ith state to be estimated. Given that the expected duration $\overline{\delta}_i$, i.e. the number of loops, in a state i is

$$\overline{\delta}_i = \frac{1}{1 - a_{ii}} \quad (10)$$

then δ_i can be proportionally obtained as follows:

$$\delta_i = \overline{\delta}_i \frac{d}{I} \quad (11)$$

3. Word graph rescoring

The ASR and the experiments reported in this work are based on the ITC-irst transcription system developed for the TC-STAR 2006 evaluation campaign [14]. The output of the system is used as the reference baseline of our approach: the exploiting of duration information is subsequently accomplished on this output as an additional, further refinement step.

The ITC-irst transcription system consists of two main components: the audio partitioner and the speech recognizer (ASR). The aim of the audio partitioner is to divide the continuous audio stream into homogeneous, non-overlapping segments and to cluster these segments into groups (*clusters*). Subsequently the ASR generates a word transcription for each cluster adopting a two pass decoding strategy. Both passes employ continuous density triphone HMMs and a trigram Language Model (LM). Conversely, the acoustic features and the acoustic normalization/adaptation procedures differ. A detailed description of the 2006 ITC-irst transcription system is reported in [14] and in the papers cited therein. The final step of the ASR module generates a Word hypothesis Graph (WG) for each incoming cluster. Each edge of a WG is associated with:

- the word transcription, w;
- initial, t^w_i, and final, t^w_f, temporal instants corresponding to the word duration $d = t^w_f - t^w_i$;
- the acoustic likelihood, $p(a|w)$;
- the language model probability, $P(w)$.

Given a WG, the corresponding best word sequence W^* is evaluated through the well known Viterbi decoding procedure:

$$W^* = \underset{W}{\text{argmax}} \sum_{w \in W} \log(p(a|w)) + \alpha \log(P(w)) \quad (12)$$

where W represents all possible word sequences inside the WG and α is the LM scaling factor. When explicit word duration probabilities are at hand, each edge can be rescored according to this additional information; the search procedure for finding the best path inside the WG has to maximize Eq. 12 modified as follows:

$$W^*, D^* = \underset{W,D}{\text{argmax}} \sum_{w,d \in W,D} [\log(p(a|w)) + \alpha \log P(w) + \beta (\log p_{LN}(d|w) - \gamma \log p_{HMM}(d|w))] \quad (13)$$

where $p_{LN}(d|w)$ is the log-normal word duration pdf, and $p_{HMM}(d|w)$ is the implicit word duration pdf explained in Section 2.2. The optimization of the weights α, β, and γ in Eq. 13 is performed over a grid of values using the development sets.

4. Experiments and results

The proposed duration model rescoring procedure was evaluated using the training, development, and evaluation data described in Table 1. These data basically consist of European (EPPS) and Spanish parliamentary speeches. For English – with a lexicon of $49k$ words – the trigram LM was trained exploiting $36M$ words of the English EPPS final text edition corpus and $200M$ words of a broadcast news corpus released by the Linguistic Data Consortium. The resulting LM was then adapted to the manual transcriptions of the EPPS audio data released for acoustic model training (about $0.7M$ words). The amount of data used to train the duration model covers 101 hours of manually transcribed speech ($1M$ of words). For Spanish – with a lexicon of $56k$ words – a trigram language model was trained on the Spanish EPPS final text edition corpus, the Spanish Parliament Texts plus the EPPS parallel corpora, about $79M$ words in total. Also in this case, the resulting LM was adapted using about $0.9M$ words from manually transcribed EPPS audio data. The explicit duration model for Spanish was trained on the manually transcribed part of the training data, i.e. 100 hours of speech that corresponds to $0.7M$ of words. The back-off strategy described in Section 2.1 for the estimation of the duration model was necessary for 10% and 24% of the Spanish and English training data respectively.

Table 2 outlines Word Error Rates (WER) obtained on EPPS tasks using the baseline system, adding the explicit duration model alone (i. e. w/o normalization: $\gamma = 0$), and introducing the two different word duration normalization methods (P^{F} and P^{P}) described in Section 2. Absolute WER reductions are about 0.2% for English and 0.1% for Spanish. There are no significant differences between the two different normalization approaches. The results reported in Table 2 show that the

	training			*dev.*	*eval.*
	AM	LM	DM		
English	176 h	$236.7M$ w	101 h	3.2 h	3.2 h
Spanish	173 h	$79.9M$ w	100 h	6.1 h	7.0 h

Table 1: Sizes [h = hours; w = words] of the data used for the experiments. Duration Model (DM) training data overlap with part of AM training data.

proposed normalization consistently leads to improvements for all LVCSR test sets in both languages.

In order to explain the limited performance gain, one has to take into account that the effectiveness of the proposed methods is upper bounded by theoretical and technical reasons. The formers include the well known trifling effect that HMM transition probabilities have on ASR performance [15]. The latters comprise WG characteristics — WG must have a low *graph error rate*, much lower than the WER that is usually the only metrics adopted for developing an ASR; models weights optimization — grid-based trimming of weighting parameters is neither optimal nor fast; heavy back-off strategies — duration models are related to words statistics and should therefore tackle the inherent data sparseness of large vocabulary tasks; intra- and inter-speaker variability — word duration densities are estimated on the whole data, independently from the speaking rate: as there are large differences between individual speakers it can be expected that the results can be further improved by adapting the duration models to the current speaker. In spite of these severe drawbacks, it is worth noticing that the proposed simple approach could, in principle, be easily extended to other types of supra-segmental features, either acoustic (e.g. pitch and energy contours) or linguistic (e.g. word-class LM).

explicit dur. model	HMM dur. normaliz.	English		Spanish	
		dev06	*eval06*	*dev06*	*eval06*
w/o	w/o	13.86	11.69	13.84	13.13
$\surd$	w/o	13.67	11.63	13.79	13.11
$\surd$	P^{F}	13.66	11.58	13.75	13.10
$\surd$	P^{P}	13.62	11.59	13.75	13.09

Table 2: Performance in Word Error Rates [%] of duration models rescoring with different HMM normalization configurations.

5. Conclusions

In this paper we have presented an approach for efficiently using word duration information during the decoding steps of an automatic speech recognition system. Word duration modeling is accomplished with log-normal probability densities, while word duration probabilities are normalized with respect to the corresponding HMM transition probabilities. Decoding is carried out in two passes: the first pass generates the word graphs that are successively rescored with word duration probabilities in the second pass. A preliminary set of experiments, carried out on LVCSR tasks ($\approx 50k$ words), shows that word durations have small effects (from 0.1% to 0.2% absolute WER reductions) on overall performance. However, we believe that the results justify further investigations, which should be devoted to exploit better strategies to combine the available information (i.e. acoustic likelihoods, LM probabilities, duration probabilities) for word graph rescoring. Contrasting experiments will be conducted to verify whether the proposed approach provides performance close to methods that use phone durations. Furthermore, we plan to investigate the unsupervised adaptation of the duration models to the speaking rate of the current speaker.

6. Acknowledgments

This work was funded in part by the European Union under the integrated project TC-STAR, Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

7. References

[1] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[2] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole, and J. Choi, "Prosody dependent speech recognition on radio news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 232–244, 2006.

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

[4] J. Pylkkonen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 385–388.

[5] B. Juang, L. Rabiner, S. Levinson, and M. Sondhi, "Recent developments in the application of hidden Markov models to speaker independent isolated word recognition," in *Proc. ICASSP*, Tampa, FL, USA, 1985.

[6] V. R. R. Gadde, "Modeling word durations," in *Proc. ICSLP*, Beijing, China, 2000, vol. 1, pp. 601–604.

[7] N. Jennequin and J.-L. Gauvain, "Lattice rescoring experiments with duration models," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006, pp. 155–158.

[8] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 307–310.

[9] J. Zheng, H. Franco, F. Weng, A. Sankar, and H. Bratt, "Word-level rate-of-speech modeling using rate-specific phones and pronunciations," in *Proc. ICASSP*, Istanbul, Turkey, 2000, vol. 3, pp. 1775–1778.

[10] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," in *Proc. ICASSP*, Atlanta, GA, USA, 1996, pp. 335–338.

[11] V. Zeissler, E. Nöth, and G. Stemmer, "Parametrische Modellierung von Dauer und Energie prosodischer Einheiten," in *Konferenz zur Verarbeitung natürlicher Sprache*, Saarbrücken, Germany, 2002, pp. 177–183.

[12] S. C. Schwartz and Y. S. Yeh, "On the distribution function and moments of power sums with lognormal components," *Bell System Technical Journal*, vol. 61, pp. 1441–1462, 1982.

[13] L. F. Fenton, "The sum of log-normal probability distributions in scatter transmission systems," *IRE Transactions on Communications Systems*, vol. 8, no. 1, pp. 57–67, 1960.

[14] F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, D. Pineda, D. Seppi, and G. Stemmer, "The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006, pp. 117–122.

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