



A Duration Modeling Technique with Incremental Speech Rate Normalization

Hiroshi Fujimura, Takashi Masuko, Mitsuyoshi Tachimori

Corporate Research and Development Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan

hiroshi4.fujimura@toshiba.co.jp

Abstract

This paper describes a novel technique to exploit duration information for low resource speech recognition systems. Using explicit duration models significantly increases computational cost due to a large search space. To avoid this problem, most of techniques using duration information adopt two-pass and N-best re-scoring approaches. Meanwhile, we propose an algorithm using word duration models with incremental speech rate normalization for the one-pass decoding approach. In the proposed technique, penalties are only added to scores of words with outlier durations, and not all words need to have duration models. Experimental results show that the proposed technique reduces up to 17% of errors on in-car digit string tasks without significant increase in computational cost.

Index Terms: speech recognition, duration model, normalized duration, decoding algorithm

1. Introduction

Hidden Markov models (HMMs) have been successfully applied to automatic speech recognition due to its flexibility and computational efficiency, though they have a lot of disadvantages. One of such disadvantages of HMMs is that they cannot model speech durations appropriately. Although the hidden semi-Markov model (HSMM)[1] which models state duration explicitly has been proposed, HSMM is not widely used for speech recognition because search space of the HSMM is significantly larger than that of the HMM.

In the past few decades, much effort has been made to model phoneme or word durations in order to improve speech recognition performance. One of the difficult problems in exploiting duration information is that it is highly affected by speech rates. To cope with this problem, one way is to estimate speech rate of input speech, and then to normalize duration or to adapt duration models. Duration modeling depending on speech rates is also included in this approach. For instance, Gadde[2] estimated speech rates for each speaker and re-scored N-best results using normalized word duration models. Wang *et al.*[3] constructed duration models for slow, normal, and fast speech rates and used them during decoding in paral-

lel. The other way is to extract duration information robust to speech rate. Ariu *et al.*[4] proposed to model duration ratio between neighboring syllables, aiming at canceling variation of speech unit duration associated with speech rate. Based on an assumption that speech rate is stable in an utterance, Zhao *et al.*[5] proposed speech rate dispersion which represents variance of speech rate of speech units in a hypothesis.

Most of the above approaches, except for [3], obtain N-best candidates and re-score them using duration information. However, for real time speech recognition systems with limited resources, algorithmic delay of N-best re-scoring is not favorable. Hence, in this paper, we propose a novel technique to apply word duration models to one-pass decoding algorithm. In our technique, speech rate is estimated incrementally, and word duration models are adapted based on the estimated speech rate.

The paper is organized as follows. Section 2 introduces the proposed duration method. Section 3 shows experimental setups and results of the proposed duration method, followed by the conclusions in section 4.

2. Word duration modeling and normalization

2.1. Word duration model based on Gamma distribution

In this paper, we use words as speech units for the duration modeling. As the word duration model, the Gamma distribution is adopted because it is reported that the Gamma distribution matches the shape of duration histogram[1].

The Gamma distribution $G(x)$ is expressed by the following equation:

$$G(x) = \frac{1}{\Gamma\left(\frac{u_w^2}{v_w}\right)} \left(\frac{u_w}{v_w}\right)^{\frac{u_w^2}{v_w}} x^{\frac{u_w^2}{v_w}-1} \exp\left(-\frac{u_w}{v_w}x\right), \quad (1)$$

where x denotes duration of word w , u_w and v_w denote mean and variance of duration of word w , respectively. By taking logarithm of the Gamma distribution,

log-likelihood $L(x)$ of the duration model is obtained by:

$$L(x) = \left(\frac{u_w^2}{v_w} - 1 \right) \log x - \frac{u_w}{v_w} x + \text{const}_w, \quad (2)$$

where

$$\text{const}_w = \frac{u_w^2}{v_w} \log \frac{u_w}{v_w} - \log \Gamma \left(\frac{u_w^2}{v_w} \right). \quad (3)$$

2.2. Normalization of word duration

In order to normalize word durations x , a normalization coefficient b is introduced. The coefficient b becomes smaller as speech rate becomes slower. Log-likelihood of the duration model of i th word w_i with the normalization coefficient b is calculated as follows:

$$L(x_i; b) = \log |b| + \left(\frac{u_{w_i}^2}{v_{w_i}} - 1 \right) \log b x_i - \frac{u_{w_i}}{v_{w_i}} b x_i + \text{const}_{w_i}, \quad (4)$$

where $|b|$ is Jacobian. Assuming that the coefficient b is constant in a hypothesis $w_{1:N}$, total log-likelihood for a hypothesis $w_{1:N}$ is obtained by:

$$L(x_{1:N}; b) = N \log |b| + \sum_{i=1}^N \left(\frac{u_{w_i}^2}{v_{w_i}} - 1 \right) \log b x_i - \sum_{i=1}^N \frac{u_{w_i}}{v_{w_i}} b x_i + \sum_{i=1}^N \text{const}_{w_i}, \quad (5)$$

As described in [6], the optimal coefficient b for the hypothesis $w_{1:N}$ in maximum likelihood sense is obtained by differentiating $L(x_{1:N}; b)$ and setting the result to zero as:

$$\frac{\partial L(x_{1:N}; b)}{\partial b} = \frac{N}{b} + \sum_{i=1}^N \left(\frac{u_{w_i}^2}{v_{w_i}} - 1 \right) \frac{1}{b} - \sum_{i=1}^N \frac{u_{w_i} x_i}{v_{w_i}} = 0, \quad (6)$$

resulting in

$$b = \frac{\sum_{i=1}^N \frac{u_{w_i}^2}{v_{w_i}}}{\sum_{i=1}^N \frac{u_{w_i} x_i}{v_{w_i}}}. \quad (7)$$

If we adopt the one-pass decoding approach, however, it is not possible to obtain all word durations in a hypothesis for a whole utterance during decoding. Hence, the coefficient b is updated word-by-word by accumulating the numerator and the denominator in Eq. (7) incrementally. The coefficient b_i for the i -th word in a hypothesis is calculated based on statistics of preceding $i - 1$ words as follows:

$$b_i = \frac{\sum_{j=1}^{i-1} \frac{u_{w_j}^2}{v_{w_j}}}{\sum_{j=1}^{i-1} \frac{u_{w_j} x_j}{v_{w_j}}}, \quad (8)$$

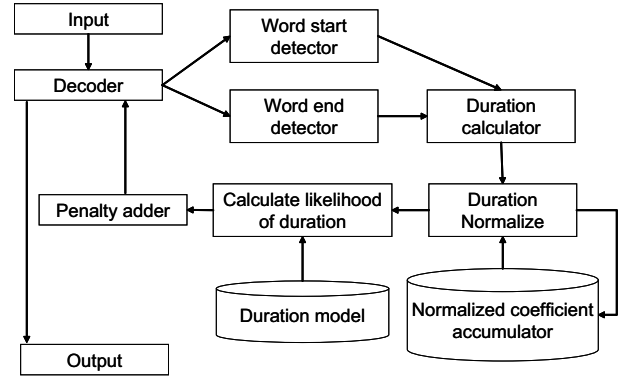


Figure 1: The block diagram of decoder using the proposed duration model.

where b_1 is set to 1. Using this normalization coefficients b_i , log-likelihood of the duration model of word w_i is calculated as follows:

$$L(x_i; b_i) = \log |b_i| + \left(\frac{u_{w_i}^2}{v_{w_i}} - 1 \right) \log b_i x_i - \frac{u_{w_i}}{v_{w_i}} b_i x_i + \text{const}_{w_i}. \quad (9)$$

Note that the normalization coefficient b_i is calculated for each hypothesis.

2.3. Decoding with Word Duration Models

Figure 1 shows a block diagram of a speech recognition system using the proposed duration modeling technique. This system has a word start detector and a word end detector which detect starting and ending points of hypothesis words, respectively. Each hypothesis holds start time of the current word and a normalization coefficient accumulators used in calculation of Eq. (8). Note that hypotheses do not have to keep the word duration history. When a hypothesis detects the word end label of the current word, duration of the word is calculated from current time and word start time held in the hypothesis. Then word duration likelihood is calculated by Eq. (9), and penalty is added to the score of the hypothesis based on the likelihood as shown in Fig.2. At the same time, accumulators of the hypothesis is updated, and the normalization coefficient is re-calculated for the next word by Eq. (8).

In this paper, penalty p_i for a word w_i is calculated from duration likelihood as follows:

$$p_i = \begin{cases} 0, & L(x_i; b_i) \geq t, \\ -c - \frac{c \cdot (L(x_i; b_i) - t)}{t}, & L(x_i; b_i) < t, \end{cases} \quad (10)$$

where c is the penalty coefficient. By setting the threshold t appropriately, penalties are added only to words with outlier duration. Furthermore, in this technique, not all

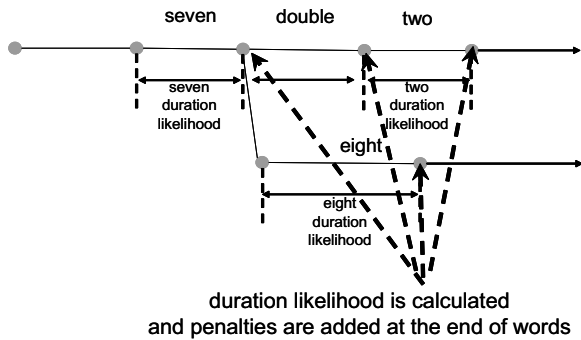


Figure 2: Decoding with word duration models.

words need to have duration models because words with appropriate duration are not given penalties. Note that zeros are added to the accumulators in Eq. (8) if a hypothesis word does not have a duration model.

2.4. Training of duration models

Word duration models are trained in an adaptive training manner as follows:

1. Initialize word duration models using the unnormalized word durations of the training data.
2. Calculate duration normalization coefficient b by Eq. (7) for each utterance in the training data using the current duration model.
3. Update word duration models using normalized word duration.
4. Repeat 2 and 3 until convergence.

3. Experiments

3.1. Experimental conditions

Experiments were conducted in three languages, i.e. American English (enUS), American Spanish (esUS), and Canadian French (frCA). Table 1 shows the number of evaluation sentences. Evaluation data were recoded in cars under idling (enon) and highway driving (hw) conditions. All sentences in the evaluation data were 10 digit strings with symbols such as ‘star’ and ‘plus.’ The numbers of utterances used for evaluation are shown in Table 1. Word loop grammars for each language were used for decoding. Only digit words had duration models, and duration models for symbols were not constructed. Word boundaries of utterances in the training data were obtained via Viterbi alignment of HMMs. Penalty coefficients c and t as well as word insertion penalties were optimized for each language.

Table 1: The number of utterances in evaluation data

	enUS	esUS	frCA
enon	1200	1695	1466
hw	1197	1678	1482

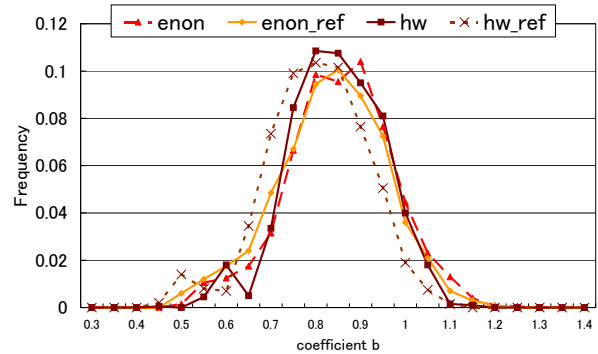


Figure 3: Histograms of Coefficient b (enUS)

3.2. Distributions of normalization coefficients

Figure 3 shows histograms of coefficient b calculated for each utterance. In this figure, “enon_ref” and “hw_ref” represent histograms of normalization coefficients calculated from correct word alignments for idling and highway driving conditions, respectively, and “enon” and “hw” represent histograms of normalization coefficients calculated at the end of decoding. Averages of “enon_ref” and “hw_ref” are 0.860 and 0.827, respectively. This means that there is a significant mismatch in speech rates between training and evaluation data; speech rates of evaluation data is slower than that of training data. Hence, it is considered that duration normalization is necessary to reduce mismatch between duration models and actual word durations in evaluation data.

Meanwhile, averages and standard deviations of differences of b for each sentence between “enon_ref” and “enon”, and “hw_ref” and “hw” were 0.027 and 0.049, and 0.025 and 0.056, respectively, while standard deviations of “enon_ref” and “hw_ref” are 0.122 and 0.117, respectively. Consequently, differences of normalization coefficients between correct sentences and recognition hypotheses are sufficiently smaller than the distributions of the coefficients, and the proposed technique is able to successfully normalize durations of hypothesis words.

3.3. Speech recognition results

Table 2 shows sentence error rates (SERs) for enUS. From this table, it can be seen that there was no improvement in performance without duration normalization. This is due to mismatch in speech rates between training and evaluation data. Meanwhile, the performance improved by normalizing duration. Consequently,

Table 2: Sentence error rates (%) for enUS

	enon	hw
Without duration	11.50	18.55
With unnormalized duration	11.33	18.63
With normalized duration	10.58	18.05

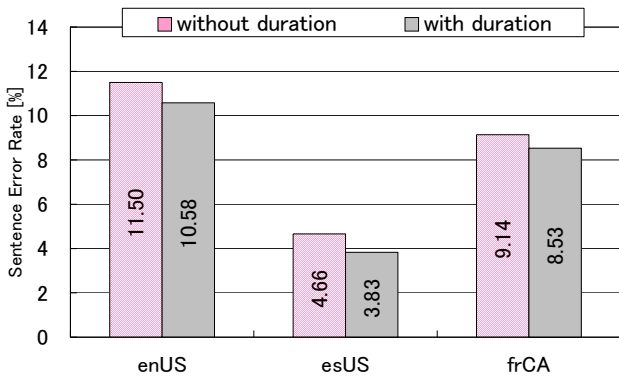


Figure 4: Sentence error rates (enon)

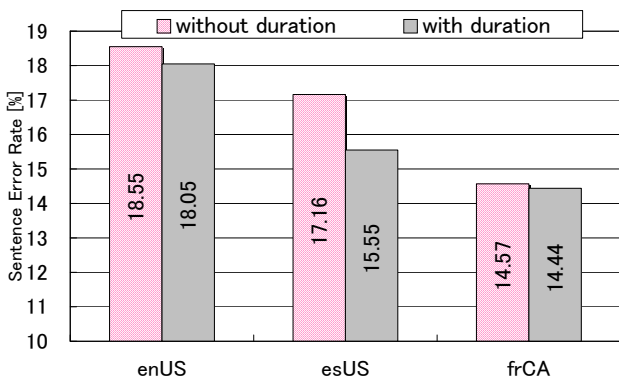


Figure 5: Sentence error rates (hw)

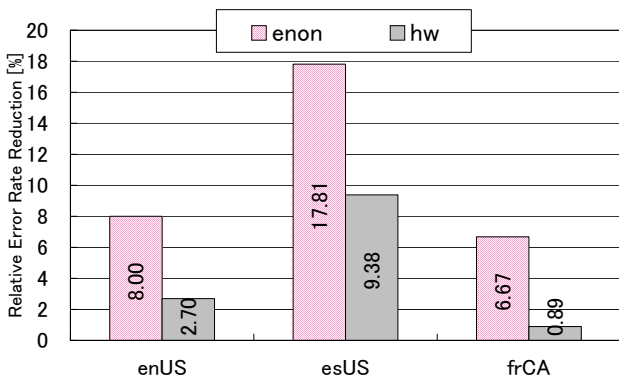


Figure 6: Relative error rate reductions

speech rate normalization is useful for duration modeling, and the proposed incremental normalization is useful in improving recognition performance.

Figures 4 and 5 show SERs without and with the proposed technique for three languages under “enon” and “hw” conditions, respectively. In this figure, “without duration” denotes SERs without duration models, “with duration” denotes SERs with normalized duration models. Figure 6 shows relative error rate reductions (RERRs) for three languages. RERRs were calculated from SERs without duration and those with normalized duration. From these figures, it can be seen that the proposed technique consistently improves performance for all languages.

3.4. Calculation time

Calculation time with the proposed technique was measured using the “time” command on FreeBSD. Evaluation data were all utterances in enUS. Relative increase in calculation time with normalized durations was only 4% compared to that without durations. Hence, the proposed technique does not significantly increase calculation time.

4. Conclusions

We proposed a new duration modeling technique with duration normalization for the one-pass decoding algorithm. In the proposed technique, normalization coefficients are updated incrementally for each hypothesis, and penalties are added only to scores of words with outlier durations. Experimental results showed that the proposed technique consistently improved recognition performance without significant increase in computational cost even if there was a mismatch in speech rate between training and test utterances.

5. 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, Mar. 1986.
- [2] V.R.R. Gadde, “Modeling word duration for better speech recognition,” *Proc. Speech Transcription Workshop*, May 2000.
- [3] W.-J. Wang and C.-J. Lee, “Duration modeling for Mandarin speech recognition using prosodic information,” *Proc. Speech Prosody 2004*, pp.591-594, Mar. 2004.
- [4] M. Ariu, T. Masuko, S. Tanaka, A. Kawamura, “Speech recognition using syllable duration ratio model,” *Proc. ICASSP2006*, vol.1, pp.1-341–1-344, May 2006.
- [5] R. Zhao, Y. Kida, X. Yan, P. Ding, L. He, “Using duration and pitch for mandarin digit string recognition,” *Proc. ICASSP2010*, pp.4846-4849, Mar. 2010.
- [6] M. Richardson, M. Hwang, A. Acero, and X.D. Huang, “Improvements on speech recognition for fast talkers,” *Proc. EUROSPEECH99*, vol.1, pp.411-414, Sept. 1999.