DURATION MODELING IN LARGE VOCABULARY SPEECH RECOGNITION

Anastasios Anastasakos† Richard Schwartz Han Shu

BBN Systems and Technologies, Cambridge, MA 02138
†Northeastern University, Boston MA 02115
E-mail: tasos@bbn.com, Tel. (617)-873-2597

ABSTRACT

This paper presents a study of different methods for phoneme duration modeling in large vocabulary speech recognition. We investigate the employment of phoneme duration and the effect of context, speaking rate and lexical stress in the duration of phoneme segments in a large vocabulary speech recognition system. The duration models are used in a postprocessing phase of BYBLOS, our baseline HMM-based recognition system, to rescore the N-Best hypotheses. We describe experiments with the 5K word ARPA Wall Street Journal (WSJ) corpus. The results show that integration of duration models that take into account context and speaking rate can improve the word accuracy of the baseline recognition system.

1. INTRODUCTION

Duration modeling has been used as a constraint in phoneme-based continuous speech recognition and modest improvements in recognition accuracy have been reported. In this paper we present an application of duration modeling in a large vocabulary continuous speech recognition system based on hidden Markov models (HMMs), the predominant technique in speech recognition in the recent years.

HMMs basically model the speech signal at the level of short time analysis frame. However, speech information is conveyed also in long term features, such as the duration of the phoneme segment. In the traditional HMM algorithms, the probability of duration of a state decreases exponentially with time. This behavior does not provide an adequate representation of the temporal structure of speech, rendering the modeling of phoneme duration weak. Consequently we find that recognition errors, arising from inadequacies in acoustic modeling, often have implausible phoneme durations. Due to this HMM inadequacy, there have been attempts to refine or replace the intrinsic duration behavior of HMMs with an explicit model at the state level, in order to penalize unlikely alignments. In this direction, hidden semi-Markov models (HSMM) allow explicit duration modeling and offer improved results [1, 2] at the cost of increasing the computation load.

Instead of improving the durational behavior at the state level, we propose to model the duration of the phonetic segment explicitly, outside the framework of HMMs after the N-best algorithm has been used to provide a list of likely hypotheses for the uttered sentence. The main advantage of the N-Best formalism [3] is that it provides the phoneme boundaries and information about the speaking rate of the candidate hypotheses. Our duration system uses this information to rescore the hypotheses and generate a score that combined with the HMM system score gives a new ranking to the sentence hypotheses.

The basic unit in measuring duration is the phoneme. The varying duration of phonemes is related to the phonetic identity, the contexts of the phoneme, the lexical stress and the speaking rate of the utterance. These features have been known to be important cues for speech perception by human listeners but not all of them have been integrated in a continuous speech recognition system. In this work, we attempt to incorporate our knowledge about the causes of the variation within the phoneme duration model.

2. CONTEXT DEPENDENCE

A natural observation is that the phonetic segment duration is affected by its neighboring phonemes. The use of different models for different contexts of the same phoneme can model the contextual effects in the durational behavior of phoneme segments. For each phoneme segment, we examined its immediate left and right context (triphone models), also taking into account the position of the word boundaries. We generated statistics for each triphone-context word-boundary dependent model, counting the number of occurences and the associated durations of each model. We smoothed these counts using the Parzen-window method to generate discrete probability density functions.

0-7803-2431-5/95 $4.00 © 1995 IEEE

The use of such detailed models is tempered by the limited amount and the uneven spread of the training data. In order to avoid poorly estimated density functions, we interpolated the triphone models with robust density functions, such as left-context and right-context (biphone models), and context independent models that are being estimated using a much larger number of occurences [4]. Since we do not want to interpolate detailed models that were estimated from a large number of training samples, the combination of the detailed and the robust model is based on the number of training samples:

$$p_{\text{final}} = \lambda \cdot p_d + (1 - \lambda) \cdot p_r \quad (1)$$

where p_d denotes the density function of the triphone word-boundary dependent model and p_r denotes the density function of the more robust model (biphone or context independent models). The weight λ is made proportional to the logarithm of the number of training tokens N_T of the triphone word-boundary dependent model:

$$\lambda = \min[1.00, 0.96 \cdot \log_{10} N_T] \quad (2)$$

3. SPEAKING RATE

One important source of variation of phoneme duration is the change in speech rate. The speaking rate is a global measure and can be defined as the average number of phonemes per unit of time. In the ARPA WSJ database, sentences are fairly long with an average of 120 phones, so a speaking rate measured over the whole sentence may ignore fluctuations within the utterance. Therefore we attribute a speaking rate measurement for each phoneme segment, which is computed based on the observed duration of a small number of phoneme segments around this phoneme. We define the speaking rate measure as:

$$r_i = \frac{\sum_{k=-M}^{M} d_{i+k}}{\sum_{k=-M}^{M} m_{i+k}} \quad (3)$$

where r_i is the speaking rate of the ith phonetic segment, d_{i+k} and m_{i+k} are the observed duration and the expected duration of the duration probability distribution, respectively, of the $(i+k)$ phonetic segment.

This definition gives us the flexibility to adjust the speaking rate computation window to any length from one phoneme segment to the whole sentence by appropriately setting the variable M. In the experiments presented, we chose to calculate the speaking rate for speech segments of five phonemes ($M = 2$), based on some preliminary experiments. Figure 1 shows the distribution of the phonetic segments occurences as a function of the speaking rate.

Given a measure of speaking rate for each phonetic segment, two different duration models were built to take

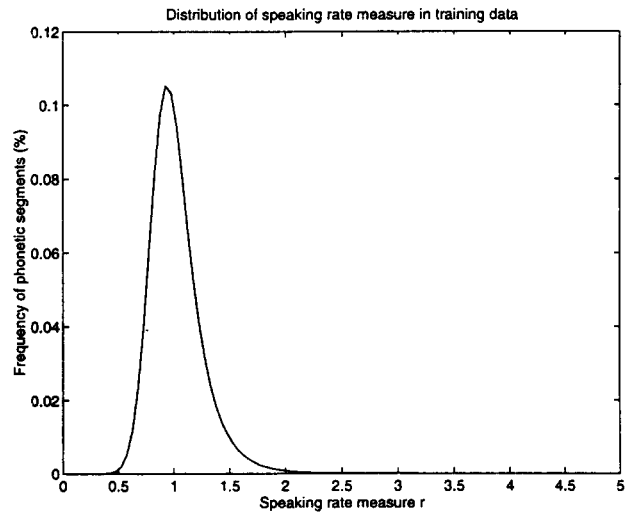

Figure 1: Smoothed histogram of the distribution of the training data with respect to the speaking rate measure r

advantage of the speaking rate information. The first approach is similar to [5] and uses the speaking rate information to cluster the training data into different sets and compute histograms and probability density functions for the duration models in each set. The second approach assumes that the duration of the phoneme segments is a function of the speaking rate and uses this assumption to generate a single normalized duration model. In both approaches, we need to generate histograms for all the duration models that occur in the training data without taking into account any speaking rate information. These statistics are necessary for the calculation of the mean duration values of the models, which is used in equation 3 in order to compute the phoneme segment.

3.1. Clustered duration models

The training phoneme segments are split into 3 groups based on the speaking rate of each segment. The partition is done so that each group contains approximately the same number of phonetic segments. Then, for each group, all duration phoneme models that occured in the group are trained according to the methods in Section 2.

3.2. Normalized duration models

In the second approach, instead of conditioning the data on speaking rate, we generated a normalized duration model. We consider duration as a function of the speaking rate. Then the normalized duration of each phonetic segment is estimated by the following linear regression equation:

$$\log \hat{d} = \alpha \cdot \log r + \beta \quad (4)$$

where $\hat{d}$ is the duration of the segment and r the corresponding speaking rate. The parameters α and β are computed, so as to minimize the estimation error in the least squares sense. We initially compute the parameters α and β from the training data and then, given the speaking rate of each phonetic segment, we compute the normalized duration $\hat{d}$ using equation 4. This way, we use all the available training data in one group to estimate a normalized duration probability density for each model. This approach provides a solution to the problem of insufficient training data that would arise, for the approach of clustered duration models, if the number of different speaking rate groups becomes large.

4. INCORPORATING LEXICAL STRESS INFORMATION

The acoustic realization of a phoneme depends also on the degree of stress, the speaker has put on it. Phonemes that do not contribute to the overall understanding of the complete utterance are likely to be poorly articulated [6]. By default the words of the lexicon that we used, are unstressed and the HMM system is trained and provides alignments using only unstressed phonemes. The simplest way to investigate the potential of stress-dependent phoneme duration modeling was to allow additional duration models for the stressed instance of a base phoneme. In this work we consider only stressed vowels. We added stress marks in the lexicon and we identified the phonetic segmentents in the alignments, provided by the HMM system, that correspond to stressed phonemes according to this stressed lexicon. We accumulated separate statistics for these instances and created separate probability densities for stressed and unstressed instances of a model. Figure 2 shows the smoothed duration distributions of phoneme segments of phoneme /AE/ with two levels of stress (unstressed and stressed segments).

5. INTEGRATING DURATION IN AN HMM-BASED SPEECH RECOGNITION SYSTEM – EXPERIMENTS

We incorporated the duration models in a post-processing phase of BYBLOS, the BBN Tied-Mixture HMM-based recognizer. The recognizer automatically labels sentence transcriptions giving the most likely word and phone alignment for each transcription. We used the labeled training data to generate the duration statistics, as described in the previous sections.

The N-Best algorithm generates, for every test utterance, a list of the N most likely hypotheses which are segmented with the HMM system. Each hypothesis is associated with the most likely word and phone alignment

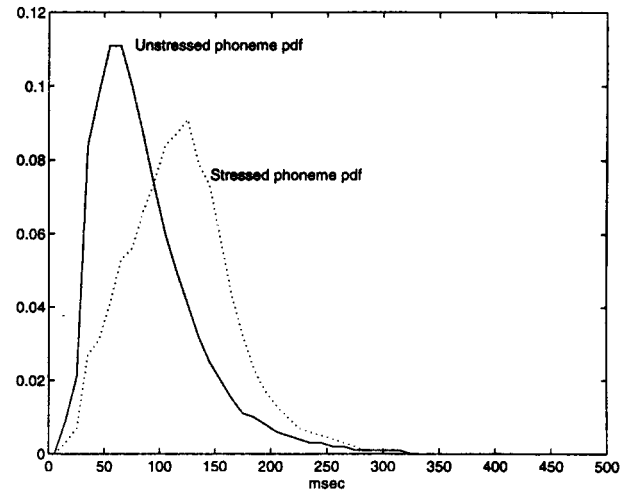

Figure 2: Smoothed duration distributions of unstressed and stressed phoneme segments of phoneme /AE/

and a log-likelihood score. Based on the segmentation of each hypothesis, the speaking rate and the duration of each phonetic segment can be determined. Then the duration likelihood of the segment can be derived using one of the duration models described in the previous sections. In the case of clustered speaking rate models, the speaking rate is used to determine a group of distributions, from which the likelihood of the duration segments is determined. In the case of the normalized duration model, the speaking rate is used to transform the observed duration to the normalized one, and based on this normalized duration we compute the duration likelihood of the phonetic segment. Alternatively we can compute the duration likelihood of a segment using a speaking rate independent duration model. Thus a duration log-likelihood score for each hypothesis is generated by summing the duration log-likelihood scores of all phonetic segments of the given alignment. The duration score can be combined with the HMM scores using appropriate weights [3] to reorder the N-Best list.

6. EXPERIMENTS-RESULTS

Experiments were carried out using the ARPA WSJ corpus [7]. This corpus consists of samples of read texts drawn from the Wall Street Journal publications and provides training and testing material for speaker independent (SI) continuous speech recognition in American English. The WSJ database is in two distinct parts – WSJ0 and WSJ1. We have built the duration system using training material from SI-37 data formed by combining training data from both WSJ0 and WSJ1 (about 36,000 utterances or 62 hours of recorded speech) and used a test set of 513

sentences having a vocabulary of 5,000 words.

The results of our experiments using duration modeling are shown in Table 1. The baseline performance of the BYBLOS system on the test set was 7.7% word error rate (WE), using a standard trigram language model. Initially the use of context independent phoneme based duration models showed an insignificant decrease in the word error rate. We achieved better results after incorporating triphone-context word-boundary dependent information reducing the error rate to 7.3%. The use of lexical stress did not show any gain. We hypothesize that the addition of lexical stress increases the number of free parameters with a possible trade-off to their robust estimation. In subsequent experiments that incorporate speaking rate information, we dismissed the lexical stress distinction of phonetic segments. The speaking rate normalized duration models did not offer any significant improvement over the speaking rate independent duration models. However adjusting the duration models to speaking rate using speaking rate clustered models reduced the error rate to 7.0%, a 10% decrease in error rate over the baseline performance. This is a significant factor considering that the ARPA WSJ database contains read speech that is expected to have little variation in speaking rate.

System	WE (%)
Baseline system	7.7%
System-A	7.6%
System-B	7.3%
System-C	7.3%
System-D	7.2%
System-E	7.0%

Table 1: Duration models used in scoring the N-Best hypotheses and percent word error (WE) performance

System-A: Context independent duration models.

System-B: Triphone context word-boundary dependent duraiton models

System-C: Triphone context word-boundary stress discriminating duration models

System-D: System-B with speaking rate normalized duration models

System-E: System-B with speaking rate clustered duration models

7. CONCLUSIONS

We modeled duration as a function of different sources of variation, such as contextual effects, speaking rate and lexical stress. We employed the N-Best algorithm, in order to add duration rescoring as a post-processing phase in an HMM-based recognizer. This approach adds very little overhead to the baseline system. We considered ways of modeling durational variations due to context by using triphone models, we proposed a local measure of speaking rate and used two different ways for modeling durational variation as a function of the speaking rate.

The results using a large vocabulary database of controlled read speech show that integration of phonetic knowledge about duration into the recognizer can improve the recognition accuracy.

8. ACKNOWLEDGMENT

This work was supported by the Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-91-C-0115, and by Ft. Huachuca under Contract Nos. DABT63-94-C-0061 and DABT63-94-C-0063 . The content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

9. REFERENCES

[1] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov model for automatic speech recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1985, pages 5-8.

[2] S. E. Levinson, " Continuously variable duration hidden Markov models for automatic speech recognition", *Computer, Speech and Language*, 1(1):29-45, 1986.

[3] R. Schwartz and S. Austin and F. Kubala and J. Makhoul and L. Nguyen and P. Placeway and G. Zavaliagkos, "New uses for the N-Best sentence hypotheses within the BYBLOS speech recognition system", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1992, pages I-1-I-4.

[4] R. Schwartz and O. Kimball and F. Kubala and M. Feng and Y. Chow and C. Barry and J. Makhoul, "Robust smoothing for discrete hidden Markov models", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1989, pages 548-551.

[5] M. Jones and P. C. Woodland, "Using relative duration in large vocabulary speech recognition", *Proceedings EUROSPEECH '93*, Berlin, Germany, September 1993, 311-314.

[6] M. Adda-Decker and G. Adda, "Experiments on stress-dependent phone modeling for continuous speech recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1992, pages I561-564.

[7] D. Paul and J. Baker, "The Design for the Wall Street Journal-based CSR Corpus", in *DARPA Speech and Language Workshop*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.