DISCRIMINATIVE DURATION MODELING FOR SPEECH RECOGNITION WITH SEGMENTAL CONDITIONAL RANDOM FIELDS

Justine T. Kao
Stanford University
Symbolic Systems Program
Stanford, CA 94309, USA
justinek@stanford.edu

Geoffrey Zweig, and Patrick Nguyen
Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{gzweig, panguyen}@microsoft.com

ABSTRACT

This paper describes a new approach to modeling duration for LVCSR using SCARF, a toolkit for speech recognition with segmental conditional random fields. We utilize SCARF's ability to integrate long-span, segment-level features to design and test duration models that help discriminate between correct and incorrect word hypotheses. We show that the duration distributions of correct and incorrect word hypotheses differ. Given a word hypothesis in the lattice and its duration, conditional length probabilities are integrated to the SCARF system as duration features. We evaluate three kinds of duration features on Broadcast News: word, pre- and post-pausal durations, and word span confusions. Adding the duration features to SCARF results in an up to 0.3% improvement over a state-of-the-art discriminatively trained baseline of 15.3% WER on a Broadcast News task.

Index Terms— duration modeling, automatic speech recognition, segmental conditional random fields

1. INTRODUCTION

Current state-of-the-art speech recognition systems e.g. [1, 2] are predominantly based on Hidden Markov Models (HMMs), and various extensions to HMMs have made them highly successful on a variety of tasks. With HMMs in such a state of refinement, duration may be one of the few aspects that are still problematic to model. The central difficulty is that HMMs in their basic form assume a fixed transition probability at each frame, resulting in an exponentially decaying distribution over individual state durations.

To address this, extensions such as hidden semi-Markov models (HSMMs) and expanded state HMMs (ESHMMs) have been proposed [3]. HSMMs model state duration explicitly using a state duration probability density function for each state of the HMM. In ESHMMs, each state is replaced by another HMM, resulting in an HMM in which the duration pdf of a given state is the overall duration pdf of the associated sub-HMM. These extensions have been shown to improve recognition performance, but mostly on isolated word recognition tests [4] or languages other than English where duration is a more prominent factor [5, 6].

Here we propose a new approach to duration modeling using Segmental Conditional Random Fields (SCRFs), as implemented by the SCARF toolkit for speech recognition [7, 8]. SCARF allows for the integration of long-span, segment-level features derived from acoustic events, and we use this ability to directly model duration distributions in a discriminative manner. Given a word hypothesis, including its duration, whether it precedes a pause, and its co-occurrence with longer or shorter words, we calculate features correlated with it being correct or incorrect. This information is then added to SCARF to help improve its recognition performance.

This paper makes two main contributions. First, we show that duration distributions have enough discriminative power to help distinguish between correct and incorrect word hypotheses. Secondly, we show that the SCRF framework is an effective way for incorporating duration features in a full-scale system. The remainder of this paper is organized as follows. In Section 2, we review the Segmental CRF framework. Section 3 presents an analysis of duration distributions on three levels: correct and incorrect hypotheses, effects of prepausal lengthening, and sets of words confused with longer or shorter words. It then shows the design and validation of duration features on these three levels. Section 4 presents integration results in the context of a complete system. Section 5 closes with a discussion of results and implications.

2. THE SCARF FRAMEWORK

As described in [7], SCARF is a toolkit for speech recognition designed to incorporate heterogeneous knowledge sources in a discriminative framework. The mathematical basis for SCARF is segmental conditional random fields (SCRFs), also known as semi-Markov CRFs [9]. Figure 1 illustrates a SCRF. The model is a two-layer model in which the top state layer represents words. The bottom layer represents observations, for example phonetic detection events. A key characteristic of SCRFs is that feature functions are defined at the segment rather than

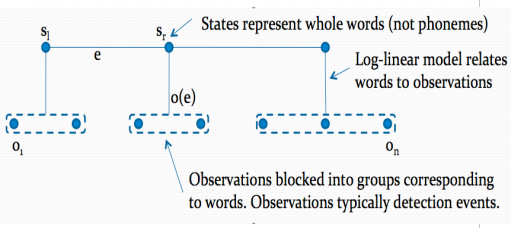

Figure 1: Model of a Segmental Conditional Random Field

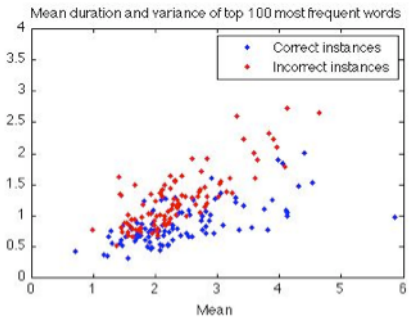

Figure 2: Mean duration and variance of top 100 most frequent words

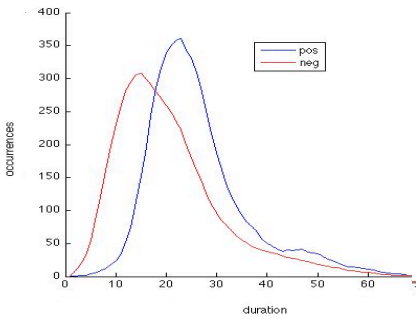

Figure 3: Word duration distributions for correct/incorrect instances of *TWO*

frame level. This is also illustrated in Figure 1, by the blocking of observations into chunks, and the assignment of one chunk to each state. When SCARF receives detections of acoustic events, a variety of features can be constructed which each measures some form of consistency between those acoustic events and a word hypothesis, and these are integrated in a log linear model.

The probability of a state sequence given the observations is given in Equation 1 below. The feature functions each involve the states s_l^e, s_r^e on the left and right hand sides of an edge *e*, and the observations associated with the edge, *o(e)*. At training the segmentation *q* of the observations into words is not known, and it is necessary to sum over all possible segmentation.

$$P(\mathbf{s}|\mathbf{o}) = \frac{\sum_{\mathbf{q}\ s.t.\ |\mathbf{q}|=|\mathbf{s}|} \exp(\sum_{e\in\mathbf{q},k} \lambda_k f_k(s_l^e, s_r^e, o(e)))}{\sum_{\mathbf{s}'} \sum_{\mathbf{q}\ s.t.\ |\mathbf{q}|=|\mathbf{s}'|} \exp(\sum_{e\in\mathbf{q},k} \lambda_k f_k(s_l'^e, s_r'^e, o(e)))}$$

To reduce computation costs, SCARF makes use of constraint lattices that represent the segmentations of observations into word sequences with appreciable probabilities modulo a baseline model. Each utterance has one numerator and one denominator constraint file. The former is restricted to paths consistent with the transcriptions, meaning it lists only correct word hypotheses, while the latter lists all the probable word hypotheses without regard to correctness. The values of user-defined features can be added to each hypothesis, which we refer to as lattice annotations. The duration features we propose are integrated into the SCARF framework as these annotations, allowing us to conveniently experiment with various forms of duration features without modifying SCARF's internal code.

For a full description of the mathematical model for SCARF, see [6, 7].

3. DISCRIMINATIVE DURATION MODELING

3.1. Duration distributions

To better understand the information available in durations, first we modeled the word duration distributions of both correct and incorrect hypotheses in the lattice. Ideally we would be able to model duration distributions for each word identity; however, most words appear too infrequently in Broadcast News to generate meaningful distributions. Thus we focused on the top 100 most frequent words that occur in the transcriptions. These 100 word identities are significant because they account for 47.5% of all word occurrences in the Broadcast News training set transcript. More importantly, they account for 48.6% of the errors in the test set. Thus, if we are able to successfully design duration features that target high-frequency words, we should be able to correct a significant portion of errors.

Normalizing for speech rate has been shown to help improve duration modeling performance [10], and so the duration of each word hypothesis in the lattice was multiplied by the phone-per-frame rate of the utterance in which it occurred. For each of the top 100 most frequent words, we calculated the (normalized) duration distributions of correct and incorrect hypotheses, and smoothed them using a 5-span moving average function. Figure 2 shows the mean and variance of the duration distributions of the top 100 most frequent words. Each data point represents a duration distribution, with the x value as the mean and the y value as the variance. Incorrect distributions (in red) tend to have higher variance and shorter durations. Figure 3 shows the plots for the duration distributions of a specific example, the word *TWO*. Blue represents the duration histogram of words correctly hypothesized as *TWO*, and red marks the histogram of words incorrectly hypothesized as *TWO* that are actually other words. For better visual comparison, the distributions are normalized so that the total numbers of occurrences for the two distributions agree. Together, these figures indicate that correct and incorrect word hypotheses do indeed differ with respect to their duration distributions.

3.2. Prepausal Lengthening

It has been shown that words preceding pauses tend to have longer durations, a phenomenon known as the prepausal lengthening effect [11]. Speech recognition systems that model pause contexts have found that it helps improve performance [10]. In order to validate and utilize the prepausal lengthening effect, we first compared the

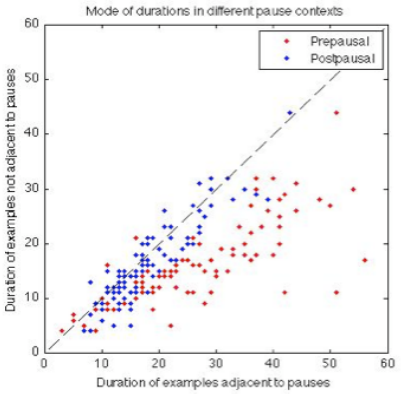

Figure 4: Durations in different pause contexts

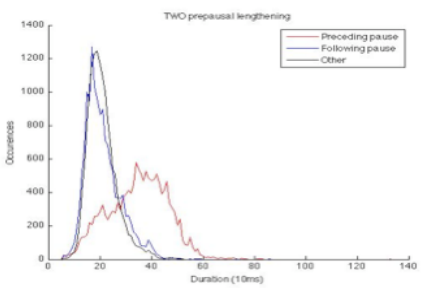

Figure 5: Prepausal lengthening example

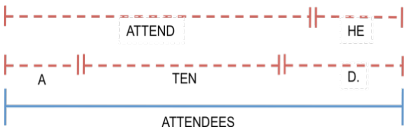

Figure 6: Word span confusion example.

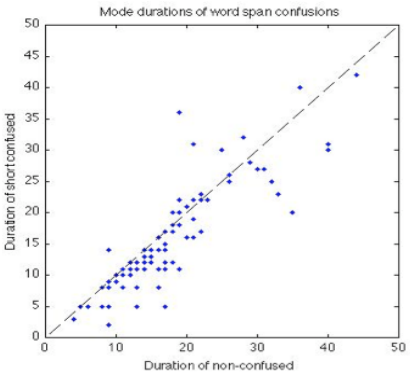

Figure 7: Durations in word span confusions

duration distributions of words preceding pauses, following pauses, and those not adjacent to pauses. Figure 4 compares the mode durations of the top 100 most frequent words in these three different pause contexts. Each data point represents a word, and the y values show the mode durations of correct instances of words that are not adjacent to pauses. The x values of the red points show mode durations of correct instances of the word when *preceding* pauses, and the x values of blue points show mode durations of correct instances of the word when *following* pauses. A word's distance from the dotted line shows the effect of pause context on its mode duration. In general, the blue dots are close to the line, indicating that durations of examples that follow pauses do not differ significantly from those not adjacent to pauses. The red dots are further away from the line, showing that words preceding pauses tend to be longer. Figure 5 illustrates the prepausal lengthening effect on the word TWO. As expected, examples of TWO that precede pauses (in red) tend to be longer than those in the other two contexts. Thus, to model the pause contexts, we should consider the durations of words preceding pauses differently.

3.3. Word Span Confusions

In addition to observing different duration distributions for correct and incorrect words and words in different pause contexts, we have observed an interesting phenomenon involving overlapping time spans. Specifically, there are cases in the constraint lattices in which a longer correct word hypothesis competes with several shorter, high frequency hypotheses that are segmentations of the longer word. We term these competing hypotheses *word span confusions.* Figure 6 illustrates one example of word span confusions. In this case, the correct hypothesis is *ATTENDEES,* and the other hypotheses are all incorrect. Such incorrect segmentations of a longer word should be more likely to have unusual durations. Figure 7 compares the mode durations of instances of top 100 most frequent words in the presence or absence of word span confusion. The x value indicates the mode duration of instances that are not confused, and the y value indicates the mode duration of examples confused with longer words. Instances confused with longer words tend to be shorter, as shown in clusters below the dotted line. This suggests that a word is less likely to be correct if there is a longer competing hypothesis whose time span completely overlaps with it. In designing duration features, we therefore consider the durations of these word span confusions separately.

3.4. Designing duration features

Based on our analysis of duration distributions, we designed three kinds of duration features: word, phone, and word span confusions.

For word duration, we introduce two features: *dur1* and *dur2*. *dur1* is defined as P(length(w) | correct) and *dur2* as P(length(w) | incorrect), that is these duration features represent the probability of the observed length given that we have a correct/incorrect instance of the word. The *dur1* and *dur2* values were calculated for all word hypotheses in the constraint lattices among the top 100 most frequent words, and set to 0 for all others. To model pausal contexts, we introduced four more features: *prepause1, prepause2, postpause1, postpause2* in addition to *dur1, dur2*. If a word hypothesis precedes a pause, then its *prepause1, prepause2* values are set to the original values of *dur1*, *dur2,* respectively, indicating that it precedes a pause, while *dur1*, *dur2, postpause1,* and *postpause2* are in turn all set to 0. If a word hypothesis follows a pause, then its *postpause1* and *postpause2* values are set to the original values of *dur1*, *dur2*, while *dur1*, *dur2, prepause1,* and *prepause2* are set to 0. Otherwise, *dur1, dur2* remain the same, and the pre- and post- pausal features are set to 0. The three sets of scores indicate the three different pause contexts we wish to model. Assigning the probabilities based on these three cases allows SCARF to learn a more fine-grained representation of the duration distributions in different contexts. Finally, for on the word span confusions, we also introduced four more features - *long1, long2, short1, short2* - alongside *dur1,*

dur2. The algorithm for assigning scores for pause contexts applies here as well. For example, if a word hypothesis is present with a shorter hypothesis within its time span, then its *long1*, *long2* values are set to the original values of *dur1*, *dur2,* while *dur1*, *dur2, short1,* and *short2* are all set to 0.

4. INTEGRATED EXPERIMENTS

To test our duration features in the context of a complete system, we chose the Broadcast News task. The experimental background is described in more detail in a companion paper [12], but briefly a 430 hour set of HUB4 and TDT4 data was used to build a state-of-the-art system using the IBM Attila system [1]. This included VTLN, LDA, MLLT, fMLLR, SAT, fMMI, mMMI, and MLLR. This system was used to create lattices to constrain the training and decoding, as well as a baseline word detector stream. To this, we added a word detector stream from a system at MSR. Altogether, with SCARF training, this gave a baseline performance of 15.3% word error rate (WER) on dev04 onto which we added the duration features. We annotated constraint lattices for the training and test sets with the duration features we designed, and trained and tested SCARF with the annotation inputs on Broadcast News. Results on dev04 are shown in Figure 8.

Duration feature(s)	WER	Absolute improvement
Baseline Attila	16.3%	-
Attila + SCARF with MSR Word Detectors (System Combination)	15.3	0.7%
Word duration	15.2%	0.1%
Word duration + Pause context scores	15.1%	0.2%
Word duration + Word span confusions	15.0%	0.3%

Figure 8: WER and absolute improvement with duration features.

Using only the word duration features *dur1* and *dur2*, the word error rate on the test set went from 15.3% to 15.2%. Adding pause context features together with *dur1* and *dur2* brought the WER down to 15.1%. Word span confusion features yielded our biggest gain. Adding word span confusion scores gave us a 0.3% decrease in WER from the SCARF baseline, resulting in a 15.0% WER.

5. DISCUSSION

The goal for adding discriminative duration features is to correct errors in SCARF's constraint lattices, and our three approaches—word, pause contexts, and word span confusions—each helped accomplish this task. It is interesting to note that word span confusion features contributed the most, suggesting that peculiarities in the lattices such as incorrect segmentations are important to consider. SCARF's flexibility with features allowed us to focus on a subset of the word hypotheses—the top 100 most frequent words—and conveniently experiment with different feature variations. Most importantly, SCARF's discriminative framework allowed us to make use of the discriminative power of durations, and model durations in a direct and practical manner.

6. REFERENCES

[1] S.F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in Speech Transcription at IBM Under the DARPA EARS Program," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006.

[2] Gales M.J.F., D.Y. Kim, Woodland P.C., Chan H.Y., R. Mrva D. and Sinha, and S.E. Tranter, "Progress in the CU-HTK Broadcast News Transcription System," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006.

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

[4] P. Ramesh and J. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Vol. I, pp. 381–384, 1992.

[5] J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[6] M. Lehr and I. Shafran. "Discriminatively Estimated Joint Acoustic, Duration and Language Model for Speech Recognition," *Proceedings of ICASSP*, 2010.

[7] G. Zweig and P. Nguyen, "SCARF: A segmental conditional random field toolkit for speech recognition," *Interspeech*, 2010.

[8] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," *Proceedings of ASRU*, 2009.

[9] S. Sarawagi and W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction," *Proceedings of NIPS*, 2005

[10] V. R. Rao Gadde. "Modeling word duration for better speech recognition," *Proceedings of NIST Speech Transcription Workshop*, 2000.

[11] W. Campbell, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, pp. 37–47, 1991.

[12] G. Zweig, P. Nguyen et al., "Speech Recognition with Segmental Conditional Random Fields: A Summary of the JHU 2010 Summer Workshop," *Submitted*