DURATION NORMALIZATION FOR IMPROVED RECOGNITION OF SPONTANEOUS AND READ SPEECH VIA MISSING FEATURE METHODS

Jon P. Nedel and Richard M. Stern

Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jnedel, rms}@cs.cmu.edu, http://www.cs.cmu.edu/~robust/

ABSTRACT

Hidden Markov Models (HMMs) are known to model the duration of sound units poorly. In this paper we present a technique to normalize the duration of each phone to overcome this weakness, with the conjecture that speech with normalized phone durations may be better modeled and discriminated using standard HMM acoustic models. Duration normalization is accomplished by dropping frames if a phone is longer than the desired duration and by adding "missing" frames and reconstructing them if a phone is shorter than the desired duration. If phone segmentations are known *a priori*, we achieve a 15.8% reduction in relative WER on spontaneous speech and a 10.3% reduction in relative WER on read speech. Preliminary work with automatic phone segmentations derived from the data is also presented.

1. INTRODUCTION

1.1. Duration normalization, speech rate, and HMMs

It is well known that HMMs do a poor job of modeling the phone durations observed in natural speech. The transition probabilities have little impact on the final hypothesis produced by modern HMM-based recognizers, and some systems have even disregarded them altogether. In 1995, Siegler and Stern reported that the duration information derived from HMM transition probabilities does not correlate well with actual duration measurements, especially when speech rate becomes more rapid or more varied [1].

There are two possible ways to alleviate this problem. One is to modify the underlying modeling structure to capture duration information more accurately. This approach might even necessitate an entirely different modeling framework. In this paper, we focus on the alternative. Our goal is to modify the data so that it is more conducive to the underlying modeling framework of choice, *i.e.* the conventional HMM acoustic models.

Our original focus was to improve recognition of spontaneous speech. We hypothesized that due to the highly irregular nature of spontaneous speech, HMMs do a poor job of capturing and modeling its characteristics. Using a parallel corpus, we warped the phone durations in spontaneous speech to match the corresponding durations observed in read speech and achieved a marginal performance improvement. This work lead to the idea of normalizing the duration of each phone in the speech data to make it more compatible with, and thus better captured and discriminated by, conventional HMMs.

Figure 1 illustrates this idea with durations abstracted from actual speech data. Continuous speech contains phones of varying duration. Each time a phone is uttered, it is produced with a different duration that depends on many different factors (*e.g.* phonetic context, speech register, speaking rate, emphasis). However, the underlying HMM that models all of the various phone renderings does a poor job of capturing duration information. Essentially, the HMM duration model is the convolution of the individual exponential duration distributions of each HMM state. This is a poor model of phone duration even if the number of states is chosen optimally for each phone. As seen in Fig. 1(a), some HMM states model a relatively short amount of speech while others are forced to model many frames of speech data with a single Gaussian mixture. Fig. 1(b) is a schematic illustration of speech that has been normalized so that every phone has the same duration. This makes the overall duration of a phone deterministic, retaining only the duration variations of the individual states within the phone. We hypothesize that duration normalization would result in reduced modeling variations across phones and improved recognition accuracy, especially for spontaneous speech where there is greater inherent variation of phone duration. This also ensures that each HMM state can capture well the specific portion of the phone it is tasked to model.

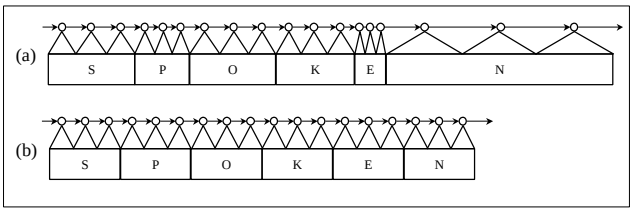

Fig. 1. Illustration of the word "spoken" before (a), and after (b) duration normalization. Corresponding HMM states are shown above each phone segment and are mapped to the approximate phone region they model.

It is also interesting to note that Stochastic Trajectory Modeling techniques imply a similar duration normalization (*e.g.* [2]).

1.2. Missing feature reconstruction methods

Missing feature methods are a popular and effective method for robust speech recognition, especially in the case of non-stationary noise (*e.g.* [3,4]). Missing feature approaches are traditionally applied to the log-spectral representation of speech. Each location in the time-frequency plane is labeled as either "present" and reliable, or "missing" and unreliable depending on the local SNR at that particular location. The "missing" or noisy speech components can then be disregarded when evaluating the overall likelihood of a given frame of speech during decoding. Alternatively, the "missing" portions can be reliably reconstructed from information contained in the clean parts of the spectrogram before the speech is decoded. In Section 2, we discuss how we use the reconstruction techniques developed for missing feature recognition to achieve phone duration normalization when phones are shorter than the desired norm.

1.3. Paper overview

In Section 2, we discuss how missing feature approaches to reconstruction of speech features can be used to normalize the duration of all phones. The experimental framework behind our experiments is detailed in Section 3. Section 4 contains experimental results demonstrating substantial improvement in recognition accuracy using normalization of phone durations for both spontaneous and clean speech assuming correct phone segment boundaries are known *a priori*. We also show some preliminary results when the phone segment boundaries are estimated blindly from the data. Finally, in Section 5 we include some discussion of our results and plans for future research that we currently are undertaking.

2. DURATION NORMALIZATION VIA MISSING FEATURE RECONSTRUCTION

In our application, we wish to normalize the duration of each phone occurrence in the speech so that every instance of a phone has the same duration. Specifically, in this paper we normalize all instances of *all* phones to have the same duration. As hypothesized earlier, this restructuring is expected to result in an improvement in performance with HMM-based modeling. The true duration of a phone can differ from the desired normalized duration: a phone can have a greater duration than what we desire (a "long phone"), or it can have a smaller duration than what we desire (a "short phone").

If a given phone segment has a greater duration than the desired normalized duration, we simply downsample the observed frame sequence. Missing-feature methods are not needed to accomplish this. However, if a phone has a duration that is less than the desired duration, we need a method for expanding its duration to the desired duration.

Missing feature methods, as discussed previously, are traditionally used to reduce the impact on recognition accuracy of unreliable time-frequency locations in the feature space that represents the speech component of the signal. In particular, time-frequency locations that are corrupted by low SNR can be *reconstructed* based on information contained in other areas of the spectrogram which are assumed to be more reliable. The same reconstruction techniques can also be used to expand and recover the "missing" portions of the phones that have a smaller duration than the desired normalized duration.

Our approach is as follows: For a given short phone, we interleave a sequence of blank frames amid the observed frames so that the new phone duration is correct. We create a missing feature mask that declares our newly-inserted blank frames as "missing" and marks them for reconstruction. The missing frames of the short phones are then filled in using the correlation-based reconstruction method described in [3].

We note that all duration normalization and reconstruction is done in the log spectral domain, in the same manner that the corresponding operation is performed for traditional missing feature reconstruction. The resulting log spectral vectors are converted to Mel-frequency cepstral coefficients for use in training and testing our standard HMM recognizer.

We had also experimented with simpler missing feature reconstruction methods, such as linear interpolation in time (which is the equivalent of simple time warping), to adjust the short phones to the correct duration. These methods resulted in no improvement in recognition accuracy. On the basis of these comparisons we believe that the added information contained in the correlations obtained from carefully-read speech allows us to regain some of the information that is lost when speech is produced very rapidly, as is often the case when speech is produced spontaneously.

3. EXPERIMENTAL FRAMEWORK

3.1. The Multiple Register Speech Corpus

We used the NIST Multiple Register Speech Corpus (MULT_REG), a parallel corpus for comparison of spontaneous and read speech recorded at SRI. The database contains fifteen spontaneous conversations on assigned topics and re-read versions of the same conversations. For our experiments, we selected data from the read and spontaneous registers. We trained and tested separate models — one for read speech and the other for spontaneous speech. We used approximately 2 hours of speech to train each acoustic model, and 0.5 hours of speech to test each model.

3.2. Speech recognizer and HMM configuration

The CMU SPHINX-III recognition system was used for all experiments. The data were modeled using 3-state left-to-right HMMs with no transitions permitted between non-adjacent states. Due to the limited amount of data in our training set, we used semi-continuous HMMs (codebook size 256). As in previous work with MULT_REG at CMU [5], the HMM states were tied based on the knowledge of phonetic context rather than on decision trees.

3.3. Log spectral correlation training

We employed a missing feature reconstruction technique that required log spectral correlations estimated from complete, clean data. The Resource Management database was used to estimate the correlations across different frequency bands at different time lags in the manner described in [3].

4. DURATION NORMALIZATION EXPERIMENTS

4.1. Experiments using oracle phone boundaries

We started by training baseline models on each of the training sets using the standard approach. In order to apply missing feature based duration normalization, we needed to know the location of the phone boundaries in both the training and the testing sets. Using the baseline models and the reference transcripts, we performed a Viterbi alignment of the transcripts to the data and derived what we deemed our "oracle" phone boundaries. Viterbi alignment was performed on both the training and testing sets used in the experiments. After alignment, however, the only information retained was the location of the boundaries that separate one phone from another.

We first focused on the data taken from the spontaneous register of the MULT_REG corpus. Given the oracle phone boundaries, we applied the missing feature methods described in Section 2 to normalize all phone occurrences in the spontaneous speech data set to a specified frame duration. We then trained standard HMM models on the duration-normalized spontaneous training set and tested their performance on the duration-normalized spontaneous test set. For the baseline WER, we also decoded the test set using the same standard models used to derive the oracle phone boundaries.

The normalized duration is a free parameter in this process; we can normalize each phone occurrence to any frame duration we choose. We empirically sought the optimal choice for the normalized duration by repeating the spontaneous speech experiment for several different normalized duration values (ranging from 6 frames to 15 frames). Note that at a normalized duration of 6 frames, the average HMM state in our 3-state models would be responsible for modeling approximately 2 frames of speech data. For a normalized duration of 9 frames, each state would be responsible for approximately 3 frames of speech data, and so forth.

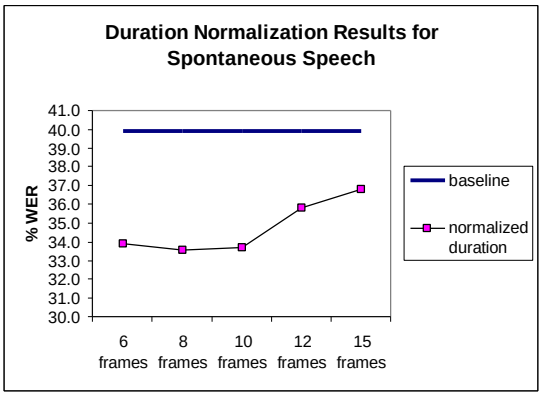

Fig. 2. Results from phone duration normalization on spontaneous speech. WER is plotted as a function of the normalized phone duration. The baseline WER is also shown for reference.

Fig. 2 plots the resulting performance of the duration-normalized models as a function of the chosen normalized frame duration. The baseline performance is plotted for reference as well. The baseline performance for the spontaneous test set was a word error rate of 39.9%. In the best case, when the speech was normalized and reconstructed so that every phone had a duration of 8 frames, the resulting WER was 33.6%. This result showed a 15.8% relative improvement over baseline performance on spontaneous speech.

Fig. 2 also shows that a choice of normalized duration in the range of 8–10 frames is best for this particular data set. When expanding to 12 or 15 frames, it is possible that correlation-based reconstruction cannot adequately estimate the missing frames. Prior experiments have indicated missing-feature reconstruction methods are only effective if the sequences of missing frames being reconstructed are no more than 5 frames long. If we expand a very short phone, say 3 frames, up to a duration of 15 frames, the missing feature methods are required to reconstruct 4 "missing" frames in a row three times in a row, with only one frame of information in between.

We then repeated the same experiment on speech taken from the read register of the MULT_REG corpus. We used the oracle phone boundaries to normalize the duration of each phone to 8 frames. We again trained standard HMMs on the duration-normalized read training set and evaluated our models on the duration-normalized read testing set. The results are shown in Table 1.

	WER	**Relative Improvement**
Baseline	15.6%	–
Normalized duration (8 frames)	14.0%	10.3%

Table 1. Results from phone duration normalization on read speech. A 10.3% relative improvement over baseline performance is shown when all phones are normalized to a duration of 8 frames.

We observed that our baseline error rate of 15.6% was reduced to 14.0% when missing feature duration normalization was applied to read speech. This reflected a relative improvement of 10.3% over baseline performance. These results show that the duration normalization methods are effective with perfect knowledge of segment boundaries, for carefully enunciated read speech as well as for spontaneous speech with its large inherent variations in phone durations.

4.2. Preliminary work for phone boundary estimation

The oracle phone segmentation experiments indicate that the missing feature duration normalization approach is promising. If the phone boundaries are known *a priori*, then a relative improvement of 10–15% in performance is possible. However, we need to be able to estimate the phone boundaries automatically for the missing-feature duration normalization method to be applicable during recognition. This is not an easy task. In this section we discuss our preliminary results in phone boundary estimation using the spontaneous speech data.

In our first simple approach, we used a blind allphone recognizer with the standard baseline models to derive the phone boundaries. This approach was not successful due to the fact that allphone recognition is not well constrained and is prone to many erroneous insertions and deletions of phone boundaries.

Our second simple approach was an iterative one. We decoded the speech using the standard baseline model. We then

Viterbi aligned the decoder hypothesis files to the speech waveforms to derive a hypothetical phone segmentation. The speech was then duration normalized using the missing feature approach as before, and decoded. The two decoder hypotheses produced up to this point were merged via a parallel hypothesis combination method reported by Singh in [6]. Specifically, the hypotheses are combined into a graph with nodes representing each word. Crossovers are introduced between the hypotheses at time instants when both hypotheses have a transition from one word to the next. (Note that if the same word is seen in both hypotheses at the same time, the two words are merged into a single node in the graph.) The graph is then searched for the best scoring hypothesis with respect to the language model.

The combined hypothesis is the output of the first iteration of the iterative decoding process, and we achieved marginal improvements over baseline performance. The process was then repeated with a Viterbi alignment of the result of the hypothesis combination process to the original data, producing a more refined phone segmentation. This led to a third hypothesis which was then combined with the other two using Singh's method to produce the final hypothesis for the second iteration. The procedure was observed to converge after three iterations and no further iterations were performed. The results after each iteration are reported in Table 2. Using the parallel hypothesis combination as an iterative approach for refining the blind phone segmentation reduced the error rate to 38.8%, a relative improvement of 2.8% over the baseline.

	WER	Relative Improvement
Baseline	39.9%	-
Iteration 1	39.2%	1.8%
Iteration 2	38.9%	2.5%
Iteration 3	38.8%	2.8%

Table 2. Results from phone duration normalization on spontaneous speech using phone segmentations derived from decoder output. Iterative incremental improvements are achieved, but we have not yet achieved the increase in performance seen when oracle phone segmentations are used.

Although we have not yet achieved the performance increases seen with the oracle phone segmentations, we are confident that by focusing directly on the problem of phone segmentation, we will be able to improve our ability to estimate the locations of phone segment boundaries.

5. DISCUSSION AND FUTURE WORK

Our experimental results indicate that duration normalization via missing feature methods is a viable approach for improved automatic speech recognition using HMM-based systems. If consistent phone segmentations are provided in the training and testing data, the approach can yield a substantial improvement in recognition accuracy for spontaneous and read speech registers. The approach yielded a 15.8% WER reduction on our challenging spontaneous speech test set. This confirms our hypothesis that recognition performance would be enhanced if the data were modified to a form that is more conducive to the underlying acoustic modeling framework.

It is clear that more work is needed in order to capitalize on the benefits of our method in the practical case when the correct phone segmentation is not known *a priori.* The segmentation problem is not a new one; and much of the research that has been done in the area of segmental modeling and landmark detection applies directly to our current problem. Although our first attempts at blind derivation of phone boundaries yielded marginal improvements in WER, they did not focus directly on the problem of segmentation. We are confident that focussed attention to the segmentation problem will yield substantial improvements in real applications. We are also working on statistical methods to reduce the dependence of the duration normalization approach on exact phone boundaries. We believe that stochastic approaches will yield further improvements in the practical system performance.

6. ACKNOWLEDGEMENTS

The authors thank Dr. Rita Singh and Dr. Bhiksha Raj for many fruitful discussions on the subject of this paper. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

7. REFERENCES

[1] M. A. Siegler, R. M. Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems", *ICASSP 1995.*

[2] Y. Gong, J. P. Haton, "Stochastic Trajectory Modeling for Speech Recognition", *ICASSP 1994*, Vol. 1, pp. 57–60.

[3] B. Raj, R. Singh, and R. M. Stern, "Inference of Missing Spectrographic Features for Robust Speech Recognition", *ICASSP 1998*.

[4] M. Cooke, P. Green, L. Josifovski, A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", *submitted for publication in Speech Communication*, 24 June 1999.

[5] J. P. Nedel, R. Singh, R. M. Stern, "Phone Transition Acoustic Modeling: Application to Speaker Independent and Spontaneous Speech Systems", *ICSLP 2000*.

[6] R. Singh, M. L. Seltzer, B. Raj, R. M. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination", *submitted for presentation at ICASSP 2001.*