Duration Normalization and Hypothesis Combination
for Improved Spontaneous Speech Recognition

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

When phone segmentations are known *a priori*, normalizing the duration of each phone has been shown to be effective in overcoming weaknesses in duration modeling of Hidden Markov Models (HMMs). While we have observed potential relative reductions in word error rate (WER) of up to 34.6% with oracle segmentation information, it has been difficult to achieve significant improvement in WER with segmentation boundaries that are estimated blindly. In this paper we present simple variants of our duration normalization algorithm, which make use of blindly-estimated segmentation boundaries to produce different recognition hypotheses for a given utterance. These hypotheses can then be combined for significant improvements in WER. With oracle segmentations, WER reductions of up to 38.5% are possible. With automatically-derived segmentations, this approach has achieved a reduction of WER of 3.9% for the Broadcast News corpus, 6.2% for the spontaneous register of the MULT_REG corpus, and 7.7% for a spontaneous corpus of connected Spanish digits collected by Telefónica de Investigación y Desarrollo.

1. Introduction

1.1. Duration normalization

In previous work [1], we had proposed an algorithm using "missing feature" methods [2,3] to normalize the duration of each phone in a speech corpus to improve the ability of the conventional acoustic Hidden Markov Models (HMMs) to discriminate among sound classes. Figure 1 is an illustration of this approach using durations abstracted from real speech data.

Each time a phone is uttered in continuous speech, it is produced with a different duration depending on many factors. As seen in Figure 1(a), the underlying HMMs contain some states which are forced to model many frames of speech data while others model a relatively short amount of speech data with the same Gaussian mixture. When speech is normalized so that every phone has the same duration, there is reduced modeling variation across phones and improved recognition accuracy, especially for spontaneous speech. The schematic in Figure 1(b) illustrates the result of duration normalization, which ensures that each HMM state can capture well the specific portion of the phone it is expected to model.

Duration normalization has been shown to be an effective approach for reducing the word error rate (WER) for spontaneous speech corpora (*e.g.* the MULT_REG corpus) when phone segmentation information is known a priori [1]. In more recent experiments with a spontaneous corpus of Spanish digits collected by Telefónica Investigación y Desarrollo, reductions in WER of 34.6% are possible when using "oracle" segmentation boundaries that are obtained by running the decoder in forced-recognition mode, aligning the output to the correct transcriptions of the utterances.
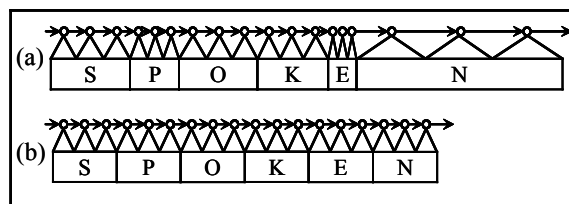

Figure 1: Illustration of the word "spoken" before (a), and after (b) duration normalization. Corresponding HMM states are shown above each phone segment and are mapped to the approximate phone region they model.

1.2. The blind segmentation problem

While duration normalization has the potential for large improvements in recognition accuracy, the problem of blind estimation of the accurate phone segmentations has continued to thwart our efforts to achieve real recognition improvements via duration normalization. Major problems occur when phone boundaries are inserted or deleted.

Figure 2 illustrates an example abstracted from our test speech data. In this example, there are two phone boundaries relatively close together in an utterance. As is often the case in spontaneous speech, there is little evidence for these boundaries in the data (probably due to phone elision), and the automatic segmentation algorithm misses these boundaries entirely. If both boundaries had been detected, the "short" segment between them would have been expanded by the duration normalization algorithm, as illustrated in the lower left of Figure 2. Because the boundaries are not detected, the little evidence for the "short" phone present in the original speech is almost completely discarded when the length of the improperly-detected "long" segment is reduced for duration normalization. This type of boundary detection error leads to a word deletion or substitution error in the final recognition hypothesis.

Similarly, when the boundary detection algorithm makes boundary insertion errors, the resulting recognition hypothesis often contains a word insertion or substitution error.

These observations led us to investigate variants of the duration normalization algorithm that would incur less devastating consequences when boundaries are missed or inserted by automatic boundary detection techniques.
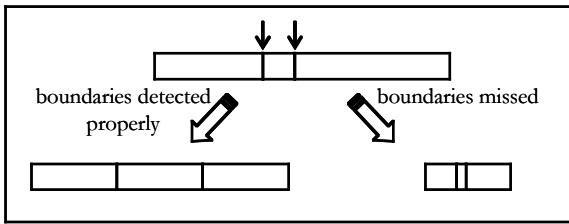

Figure 2: Illustration of resulting normalized segments when boundary detection is in issue. If the indicated boundaries are not detected, three actual phone segments in the original speech are presumed to be a single "long" phone segment. When the length of this improperly-labeled "long" segment is reduced for duration normalization, most of the information from the darkened segment is lost.

1.3. Paper overview

In Section 2 we discuss three different variants of the duration normalization algorithm that we investigated. Each variant independently generates a recognition hypothesis for a given speech utterance. These hypotheses are then combined to generate a final hypothesis. Section 3 details our experimental setup. In Section 4 we report experimental results showing significant recognition performance improvements when duration normalization and hypothesis combination are used in conjunction. These results are obtained using automatically-derived segmentation information. Section 5 contains a discussion of the results and future work.

2. Variants of duration normalization

For a given phone segment, the duration normalization algorithm will do one of two things. If the phone is longer than the desired duration (a "long" phone), the sequence of log-spectral vectors corresponding to the phone is downsampled in time to achieve the normalized duration. If the phone is shorter than the desired duration (a "short" phone), the log-spectral vectors are expanded in time, and the "missing" vectors are replaced by missing feature methods [2].

As described in Section 1.2, boundary detection errors often lead to recognition errors, especially in cases when short phones are not detected. To alleviate this problem, we experimented with the following variants of duration normalization:

Standard: expand short phones, contract long phones

Expand-only: expand short phones, leave long phones at their natural duration

Contract-only: contract long phones, leave short phones at their natural duration

The expand-only variant helps to compensate for examples like the illustration in Figure 2. If the boundaries of a "short" phone are missed, the surrounding segment would be incorrectly considered a long phone and contracted by the standard duration normalization approach.

In expand-only duration normalization, the incorrect long phone would not be contracted in time, giving us a better chance to properly recognize the missed short phone during decoding. Similarly, contract-only duration normalization helps to compensate for spurious boundaries inserted by automatic boundary estimation algorithms.

Each variant of duration normalization gives rise to a different set of acoustic models during training and a different recognition hypothesis during decoding. By design, decoding with expand-only duration normalization should produce fewer word deletion errors but more word insertion errors. Conversely, decoding with contract-only duration normalization should result in more word deletion errors and fewer word insertion errors. These systematic variations should make the hypotheses good candidates for merging via the parallel hypothesis combination method reported by Singh in [4].

In Singh's method, 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.

3. Experimental framework

3.1. Speech corpora

We used three speech databases for our experiments:

TID: A database of spontaneous Spanish connected digit strings collected by Telefónica Investigación y Desarrollo in Madrid, Spain. The data were collected from cellular telephone calls in which speakers were asked to remember and repeat connected digit strings and monetary amounts. The data set contains approximately 7 hours of training speech and 2.5 hours of testing speech data.

MULT_REG: NIST Multiple Register Speech Corpus, a parallel corpus for comparison of spontaneous and read speech recorded at SRI International. 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 spontaneous register. Our training set contains approximately 3 hours of training data and 1 hour of testing data.

BN: NIST Broadcast News data taken from previous HUB 4 evaluations. Models were trained on 45 hours of speech taken from the 1996 and 1997 corpora. Testing was done on the 1999 Eval 1 data set.

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. For TID and MULT_REG, we used semi-continuous HMMs (codebook size 256) due to the limited amount of data in our training sets. For BN, we used fully continuous HMMs with a mixture of 16 Gaussians per state.

4. Experiments

We started by training baseline models for each of the training sets using standard techniques. Duration normalization requires knowledge of the location of the phone boundaries in both the training and the testing sets. In our "oracle" experiments, we used the baseline models and the reference transcripts and performed a forced Viterbi alignment of the transcripts to the data to derive "oracle" phone boundaries. In our "blind" experiments, we decoded the speech using the baseline models and aligned the resulting recognition hypotheses to the data to derive the locations of our estimated phone boundaries.

Using these phone boundaries, we then normalized our training and testing sets using each of the three variants of duration normalization (standard, expand-only, contract-only). For each corpus, we trained three separate acoustic models on the training set, one model for each variant of duration normalization.

We then decoded the testing sets using each variant of duration normalization, which produced three recognition hypotheses for a given utterance. Finally, we employed hypothesis combination [4] to select the final recognition hypothesis and scored our results. Table 1 reports results for TID data. Table 2 contains results for MULT_REG data. BN results are reported in Table 3.

TID results	WER	Relative Improvement
Baseline	5.2%	—
"Oracle" experiment	3.2%	38.5%
"Blind" experiment	4.8%	7.7%

Table 1: Results for duration normalization and hypothesis combination on the TID Spanish connected digits data. This technique achieves a 7.7% relative reduction in WER on TID data.

MULT_REG results	WER	Relative Improvement
Baseline	40.3%	—
"Oracle" experiment	31.7%	21.3%
"Blind" experiment	37.8%	6.2%

Table 2: Duration normalization and hypothesis combination results for the spontaneous register of the MULT_REG corpus. A relative reduction in WER of 6.2% is achieved on MULT_REG data.

BN results	WER	Relative Improvement
Baseline	33.4%	—
"Oracle" experiment	28.8%	13.8%
"Blind" experiment	32.1%	3.9%

Table 3: Broadcast News 1999 Eval 1 recognition results with duration normalization and hypothesis combination. A relative reduction in WER of 3.9% is achieved on BN data.

Our experimental results show a reduction in WER over baseline for each of the databases tested. Consistent with experiments using various speech compensation algorithms for robust recognition, the performance improvement achieved using smaller databases is greater than the performance improvement achieved using larger databases such as Broadcast News. We believe that this occurs because in large tasks the extensive amount of training data and detailed modeling framework lead to a system that is inherently more robust. Nevertheless, the use of duration normalization in conjunction with hypothesis combination algorithm yields a performance improvement even in the large-scale BN test.

We note that when using standard duration normalization alone with oracle segmentations, the best possible reduction in WER is 34.6% for TID, 20.1% for MULT_REG, and 5.4% for BN. Standard duration normalization alone with estimated segmentations does not yield significant improvements over baseline performance on any of the databases tested. When duration normalization is combined with hypothesis combination, significant improvements are achieved in all of our tests.

4.1. Error analysis for variants of duration normalization

Table 4 shows the breakdown of errors made by each variant of duration normalization using estimated segmentation information on the MULT_REG corpus. The word recognition errors are broken down into substitution (sub.), deletion (del.), and insertion (ins.) errors. The baseline error breakdown and post-hypothesis combination error breakdowns are also given for reference.

As expected, expand-only duration normalization produces fewer word deletion errors and more word insertion errors than standard duration normalization. Also, contract-only duration normalization produces fewer word insertion errors and more word insertion errors than standard duration normalization. Hypothesis combination is able to take advantage of these variations to produce recognition hypotheses with a lower word substitution rate than any of the single duration normalization variants alone. Note that similar trends are observed with TID and BN data as well.

MULT_REG WER breakdown	Sub. errors	Del. errors	Ins. errors
Baseline	23.2%	11.9%	5.2%
Standard dur. norm.	22.2%	13.7%	3.9%
Expand-only dur. norm.	23.0%	12.8%	4.5%
Contract-only dur. norm.	22.1%	13.9%	3.6%
Dur.norm. + hyp. comb.	20.7%	13.6%	3.5%

Table 4: Types of recognition errors made by each variant of duration normalization with estimated segmentation information on MULT_REG data. Word recognition errors are broken down into substitution (sub.), deletion (del.), and insertion (ins.) errors.

Table 5 shows a complete result summary for each variant of duration normalization applied to the MULT_REG corpus. Again, baseline and post-hypothesis combination results are also given.

Using blindly-estimated segmentation information, slight improvements are made by the standard and contract-only variants of duration normalization alone. Although the expand-only variant makes different types of errors from the baseline, it does not reduce the overall word error rate. Hypothesis combination of the recognition output produced by the duration normalization variants outperforms the individual hypotheses produced by each variant alone. As stated earlier, when duration normalization and hypothesis combination are used in conjunction on the MULT_REG corpus, a 6.2% relative improvement in performance over baseline WER is achieved.

MULT_REG result summary	WER	Relative Improve-ment
Baseline	40.3%	—
Standard dur. norm.	39.8%	1.2%
Expand-only dur. norm.	40.3%	0%
Contract-only dur. norm.	39.6%	1.7%
Dur. norm. + hyp. comb.	37.8%	6.2%

Table 5: Summary of errors made using duration normalization and estimated segmentation information on the MULT_REG corpus. Hypothesis combination of the individual recognition hypotheses produces a 6.2% relative performance improvement over the baseline.

It is interesting to note that with TID and BN data, none of the individual variants of duration normalization alone produces an improvement in recognition performance. For BN, we observe that the overall performance is actually worse than baseline for each of the three variants of duration normalization alone. However, because our approach is designed so that each variant makes different types of errors, we are able to achieve improvements in recognition performance in spite of the fact that the individual hypotheses produce worse WERs than baseline.

5. Discussion and future work

Our results show that duration normalization is a practical technique for improving speech recognition performance for HMM-based systems when the recognition hypotheses produced by its variants are combined with hypothesis combination.

Further examination of our results confirms that expand-only duration normalization produces recognition hypotheses with a higher word insertion rate and a lower word deletion rate than the other variants of duration normalization. Also, the recognition hypotheses generated by contract-only duration normalization have a lower word insertion rate and a higher word deletion rate than the other variants. Hypothesis combination is a successful method to combine these individual hypotheses and choose a good overall hypothesis.

When duration normalization is combined with hypothesis combination, there is a greater improvement in recognition performance than with duration normalization alone. With oracle segmentations, we see a greater potential for improvement than that of standard duration normalization alone. With estimated segmentations and standard duration normalization, we generally do not observe improvements in recognition performance. With estimated segmentations, duration normalization, and hypothesis combination, we achieve significant improvements in recognition performance on all databases tested, including a more rigorous experiment on a large vocabulary Broadcast News recognition task.

Future work will investigate alternate methods for combining the individual recognition hypotheses produced by the duration normalization variants. In [5], Li reports a technique to combine the word lattices produced by different recognition systems to find an optimal recognition hypothesis. Li's lattice combination technique consistently outperforms Singh's hypothesis combination technique. We plan to investigate the effectiveness duration normalization when combined with lattice combination.

We also plan to investigate "soft" variants of duration normalization which can make use of probabilistic rather than hard phone boundary decisions.

6. Acknowledgements

7. References

[1] J. P. Nedel and R. M. Stern, "Duration Normalization for Improved Recognition of Spontaneous and Read Speech via Missing Feature Methods", *Proc. ICASSP* 2001.

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

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", *Speech Communication*, **34** (3): 267–285 (2001).

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

[5] X. Li, R. Singh, and R. M. Stern, "Lattice Combination for Improved Speech Recognition", *Proc. ICSLP* 2002.