Modeling Between-Word Coarticulation in Continuous Speech Recognition

Mei-Yuh Hwang, Hsiao-Wuen Hon, Kai-Fu Lee

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

This paper describes the addition of between-word coarticulation modeling into SPHINX, an accurate large-vocabulary speaker-independent continuous speech recognition system. Between-word coarticulation is a major source of phonetic variability in continuous speech. By detailed modeling of between-word triphones and utilizing the generalized triphone technique, we obtain an error rate reduction of 16% to 29% for different test sets and grammars on the DARPA Resource Management (RM) Task.

1. Introduction

Context-dependent phoneme models, such as triphones [1] or generalized triphones [2], have become a very practical and successful class of subword units in large-vocabulary continuous speech recognition. These phoneme-sized models take into account the neighboring phonetic contexts, which strongly affect the realization of a phoneme. However, previous approaches have only considered within-word coarticulation and have ignored between-word coarticulation, which is especially important in continuous speech. This paper describes the addition of between-word coarticulation modeling into the SPHINX system [3], which is based on phonetic discrete hidden Markov models (HMMs).

A simple extension of triphones to model between-word coarticulation is problematic because the number of triphone models grows sharply when between-word triphones (BWTs) are considered. For example, there are 2381 within-word triphones (WWTs) in our 997-word task while there are 7549 triphones in our training data when BWTs are also counted. For SPHINX, 7549 models mean over 17,000,000 parameters, which we cannot hope to train with the limited training set.

Therefore, generalized triphone models [3] are used to combine similar triphone contexts, and thus to reduce the number of models. These 7549 triphones are categorized into 1100 generalized triphones by an automatic clustering procedure. These 1100 models are then tested in the recognizer. The main modifications to incorporate BWTs lie in the construction of sentence models in the trainer and the word transitions in the search space of the recognizer.

With BWTs, the recognition error rates of the DARPA RM Task are reduced by 16% to 29% on various test sets.

2. Speech Representation

The speech is sampled at 16 KHz, and pre-emphasized with a filter of \(1 - 0.97t^{-1}\). Then, a Hamming window with a width of 20 msec is applied every 10 msec. Autocorrelation analysis with order 14 is followed by LPC analysis with order 14. Finally, 12 LPC-derived cepstral coefficients are computed from the LPC coefficients, and these LPC cepstral coefficients are transformed to a mel-scale using a bilinear transform.

These 12 coefficients are vector quantized into a codebook of 256 prototype vectors. In order to incorporate additional speech parameters, we created two additional codebooks. One codebook is vector quantized from differential coefficients. The differential coefficient of frame \(n\) is the difference between the coefficient of frame \(n+2\) and frame \(n-2\). This 40 msec. difference captures the slope of the spectral envelope. The other codebook is vector quantized from energy and differential energy values. The use of multiple codebooks was introduced by Gupta, et al. [4].

3. Four Types of triphones

When adding BWTs, we also notice the phonetic variability of the same triphone but at different word positions (beginning, middle, or ending). In order to take into account all the phonetic variability in WWTs, BWTs, and phonetic locations, we come up with 4 types of triphones. The first type is a triphone embedded within a word, such as the /t/ phone in THEATRICAL. We will use the notation \(P(L, R)\) to represent a triphone \(P\) with a left context phone \(L\) and a right context phone \(R\). \(SIL\) represents the model of silence. Hence, the /t/ in THEATRICAL will be written as \(\tau(AE, ROCK)\).

The second type of triphone appears at the beginning of a word and is postfixed by the letter \(b\). For example, the \(R(T, PA)\) b phone in THAT ROCK. Similarly, the third type of triphone appears at the end of a word and is appended with the letter \(e\), like the \(\tau(AE, R)\) e phone in THAT ROCK. Note even though the /t/ phones in THEATRICAL and THAT ROCK have the same left and right contexts, they are modeled by two different triphones because of their different word positions.
Short words are affected the most by between-word coarticulation. This is especially true for single-phone words. Therefore, we think it may be a good idea to model single-phone words separately. This type of triphone is postfixed by the letter s. For example, the AX (z, w) s phone in THAT IS A WORD. With these 4 types of triphones, the total number of triphones comes to 7549.

4. The Trainer and Sentence HMMs

Figure 1 shows the training procedure in SPHINX. For each input sentence text, the trainer first builds a corresponding sentence HMM, using context-dependent phone models as units. Then it runs the forward-backward algorithm [5] on the sentence HMM using the associated input speech. After running forward-backward on the whole set of training sentences, it reestimates the HMM parameters using the Baum-Welch formulas. The entire reestimation process is iterated several times to get better trained parameters. To recover from poor estimation of some parameters, the context-dependent parameters are interpolated with context-independent ones [6].

For double-phone and single-phone words, word-boundary connections are a little more complicated as shown in Figure 3 and Figure 4. For double-phone words, the two beginning phones must be merged before the two ending phones branch. On the other hand, among the four triphones of a single-phone word, each pair of them with the same right context must be merged before going on the next word.

5. Generalized Triphones

As discussed in earlier sections, 7549 triphones means over 17,000,000 parameters. With only 4000 training sentences or so, most parameters will be zeroes. Therefore we need some way to reduce the number of HMMs. The same dustering procedure introduced in [3] is used to merge similar triphones together. The outline of the clustering procedure is as follows:
1. An HMM is generated for every triphone context.

2. Clusters of triphones are created; initially, each cluster consists of one triphone.

3. Find the most similar pair of clusters which represent the same monophone, and merge them together.

4. For each pair of clusters, consider moving every element from one to the other.
   - Move the element if the resulting configuration is an improvement.
   - Repeat until no such moves are left.

5. Until some convergence criterion is met, go to step 3.

To determine the distance between two models, we use the following distance metric:

\[
D(a,b) = \frac{\sum_{i} (P_{a}(i))^{N_{a}(i)} \cdot (P_{b}(i))^{N_{b}(i)}}{\sum_{i} (P_{m}(i))^{N_{m}(i)}},
\]

where \(D(a,b)\) is the distance between two models of the same phone in context \(a\) and \(b\). \(P_{a}(i)\) is the output probability of codeword \(i\) in model \(a\), and \(N_{a}(i)\) is the count of codeword \(i\) in model \(a\). \(m\) is the merged model by adding \(N_{a}\) and \(N_{b}\). In measuring the distance between the two models, we only consider the output probabilities, and ignore the transition probabilities, which are of secondary importance.

Thus, we first train the whole 7549 BWTs and WWTs. Then after clustering, we train these generalized triphones using the same training procedure but with triphones mapped into corresponding generalized triphones. To obtain more robust models, we interpolate these trained generalized triphones with their context-independent phones. Our best system currently has 1100 generalized triphones. The WWT system mentioned in Section 7 contains 1076 models.

6. The Recognizer and the Language Network

The recognizer runs the Viterbi beam search [1, 8] to find the sequence of states that maximizes the probability of generating the input speech given the models obtained by the trainer. The sequence of words backtraced from the sequence of states is output as the recognized sentence.

The search space of the recognizer, which we call the language network, is the set of vocabulary plus connections between words. In SPHINX, the language network is pre-built though it can also be generated dynamically in the recognizer. In this network, word models are formed in the way similar to that in the trainer. However, each word now has multiple beginning and ending triphones since there are more than one word that can precede or follow it. Also words are connected in two ways (one through pause, the other without pause) by appropriate BWTs. For example, in Figure 5, since the ending TS (AH, SIL) e of WHAT+S has a right context of SIL, it is connected to SIL. Similarly, TS (AH, AX) e is connected to AX (TS, M) s and AX (TS, SIL) s of the word L, and TS (AH, DH) e to DH (TS, AX) b of the word THE. Note that although triphones are used in Figure 5, the models we use in the recognizer are generalized triphones.

One possible problem for the BWT language network is that the 7549 triphones are constructed from the training data. Due to the incompleteness of training sentences, there exist other BWTs which are allowed by the grammar but are not observed in the training data. When a triphone is missed in the language network, we use its context-independent phone instead.**

**We have also done experiments with Word-Context-Free phones as Lincoln Lab does. It yields almost the same accuracy as the context-independent phones.
7. Results

The above between-word coarticulation modeling is applied to the 997-word DARPA RM Task, both with and without the word-pair grammar, using 3990 training sentences (109 speakers) and two test sets: June88 with 300 sentences (12 speakers) and Feb89 with another 300 sentences (12 speakers). With the word-pair grammar, which knows only about the legality of pairs of words, the perplexity is about 60. Without it, the perplexity is 997. The word accuracies of the WWT and BWT systems on the two test sets with and without the word-pair grammar are shown in Table 1 and 2. Word accuracy is defined as the percent of words correct minus the percent of insertions.

<table>
<thead>
<tr>
<th>Models</th>
<th>WWT</th>
<th>+ BWT</th>
<th>% Error</th>
<th>Reduct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1076</td>
<td>1100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WPair</td>
<td>90.1%</td>
<td>92.6%</td>
<td>17.8%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: SPHINX results on June88.

<table>
<thead>
<tr>
<th>Models</th>
<th>WWT</th>
<th>+ BWT</th>
<th>% Error</th>
<th>Reduct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>66.2%</td>
<td>71.8%</td>
<td>16.6%</td>
<td></td>
</tr>
<tr>
<td>WPair</td>
<td>91.0%</td>
<td>91.2%</td>
<td>17.8%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: SPHINX results on Feb89.

During our experiments, we found that the usual 2 iterations of reestimation is not quite enough. It is probably because the learning problem is harder and takes a little longer to learn. The substantial error rate reductions show that modeling of between-word coarticulation is indeed worthwhile.

8. Conclusion

We believe that the choice of a speech unit represented by an HMM is crucial. Previous studies have proved that triphone models are a practical and successful class of units. They model coarticulation effects caused by neighboring phonemes. However, most studies have ignored between-word coarticulation, which is especially important in continuous speech. In order to take it into account, we add between-word triphone models into the SPHINX system. Not only adding BWTs, we also consider speech variability at different word positions. The considerations in WWTs, BWTs and phonetic locations give us four types of triphones and sum up to 7549 triphones. However, to control the growth of the number of parameters to be estimated, we use generalized triphones instead of the large number of complete triphones directly. We first train the 7549 triphones and then merge them into 1100 clusters according to their similarity. These 1100 generalized triphones are then trained and used in the recognizer. We report 16.6% to 29.3% error reduction by adding between-word triphones only.

This work has shown that while HMM learning is very powerful, it is essential to use detailed models guided by speech knowledge, and to apply principled techniques to control the growth of the parameters. In the future, we will explore the modeling of other causes of phonetic variability.

Acknowledgments

The authors wish to thank Raj Reddy for his encouragement and support. This work was sponsored by DARPA.

References