PROSODY AS A CONDITIONING VARIABLE IN SPEECH RECOGNITION

Sarah Borys[1], Mark Hasegawa-Johnson[1], and Jennifer Cole[2]

1. Department of Electrical and Computer Engineering
2. Department of Linguistics
University of Illinois at Urbana-Champaign, Urbana, IL 61901

ABSTRACT

In this paper, we demonstrate two different methods for improving the accuracy and correctness of the standard HMM speech recognizer. The first method involves incorporating prosody into the acoustic model. The second method modifies the HMM so that it can make use of explicit duration probability densities to improve recognition rates. We prove that both of these methods are effective by running two different experiments using the Hidden Markov Toolkit (HTK). The first experiment determines the correctness and accuracy of a model by performing word recognition. The second experiment determines how well a model picks up on the prosody.

1. INTRODUCTION

A spoken sentence in the English language can be separated into smaller phrases based on the speeding up and slowing down of words in 4 or 5-word groups. The short pause or slowing down at the beginning and end of each of these phrases is what is referred to as a phrase boundary. These phrase boundaries have important effects on spoken words. Previous studies have indicated that different acoustic phenomena occur near phrase boundaries. One such phenomenon is the increase or decrease in phoneme duration near the intonational phrase boundary. Wightman [1] reports that there is segmental lengthening in the rhyme of the final syllable of a pre boundary word. Fougeron and Keating [2] discovered that final vowels and initial consonants of boundary words have less reduced lingual articulations compared to vowels and consonants that occur elsewhere in the phrase. Along with this change in articulation, Fougeron and Keating also report that lengthening tends to occur near prosodic boundaries. These findings suggest that there is information in the final and initial syllables of boundary words that can be made use of in automatic speech recognition (ASR). The information contained in phrase boundaries is just one usable type of prosodic information. Prosody (described in detail in section 3) is a relatively new concept in ASR and to our knowledge, very few people have done experiments which directly incorporate prosody with ASR. Kompe [3] is one of the few people to experiment with prosody. He reports improvements to recognition rates when prosodic information is used for recognition purposes. Chen, Borys and Hasegawa-Johnson [4] preformed experiments that used prosodic information along with explicit durations and found that systems that use prosody show improvement over systems that do not. This paper extends on Chen, Borys, and Hasegawa-Johnson's work by performing additional experiments using the same training and testing techniques that were used in [4].

2. HIDDEN MARKOV MODELS

Most modern speech recognizers use hidden Markov models (HMMs) to perform recognition. In this section, we will provide a simplified explanation of HMMs based on the explanation in [5].

An HMM can best be thought of in terms of a state diagram. The example diagram in figure 1 is made up of three states. Each individual state is associated with what are called transition probabilities. Transition probabilities determine how likely it is that a state will be exited at time t+1. In state i, there is also a function $b_i(x)$ that gives the probability density of observing vector x if the HMM is in state i. The idea behind a HMM is that the number of states and all the probabilities (transition and event) are known during the recognition process (they have been learned during the training process), but the actual state that the model is in at any given time t is unknown. This is where the term "hidden" comes from.

As an example, suppose the states in figure 1 were actually baskets of different colored gumballs. For simplicity, we'll say that there are only two colors of gumballs. The observation we will make in each state is the color of gumball that is randomly drawn out of the current basket. Each basket has different amounts of both colors, so the probability of

observing one color over the other would depend on the basket we are choosing from. After Δt seconds, we will be handed a new basket based on the transition probabilities associated with the basket we just drew from. It is possible that we could end up choosing from the same basket after we make a transition, but we will never actually know which basket we are choosing from. This process will repeat N times to give us an observation sequence O.

There are two major concerns with HMMs that need to be addressed. First, there is the recognition problem; how do we calculate $Pr(O|\lambda)$, the probability of observation sequence $O = (O_1, O_2, O_3, .., O_T)$ given the model, λ? Secondly, there is the training problem; how can the model learn the probabilities (transition and event) in order to best model some training example O.

The most efficient method to calculate the probability of an observation sequence O for any given model is called the forward-backward method. This method involves the calculation of the forward variable, $\alpha_t(i)$, and the backward variable, $\beta_t(i)$. To calculate these
variables, let us assume that we will be making observations for a total time of T seconds.

The forward variable, $\alpha_t(i)$, is defined as the probability of a partial sequence of observations and states up to time t, given the current model. $\alpha_t(i)$ can be calculated as follows:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}),$$

where

$$t = 1, 2, .., T-1, \ 1 \le j \le N$$

where N is the number of states in the model, π_i is the initial state distribution, O_t is the observation at time t, b_j is the observation symbol probability distribution in state i, and a_{ij} is the state transition probability distribution from state i to state j. We can now calculate $Pr(O|\lambda)$ with the equation:

$$Pr(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

The backward variable $\beta_t(i)$ represents the probability that a partial observation sequence will be observed from time t+1 until time T given states q_i at time t. $\beta_t(i)$ can be calculated using the following method:

$$\beta_T(i) = 1$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}(j))$$

where

$$t = T-1, T-2, .. , 1, \ 1 \le i \le N$$

The forward variable $\alpha_t(i)$, and the backward variable, $\beta_t(i)$ can now be used to calculate $\gamma(i)$, the probability of being in state q_i at time t.

$$\gamma(i) = \frac{\alpha_t(i)\beta_t(i)}{Pr(O|\lambda)}$$

To determine the most likely state, i_t, at time t, we take:

$$i_t = \text{argmax}[\gamma_t(i)]$$

where

$$1 \leq t \leq T,\ 1 \leq i \leq N$$

The training problem is best solved as follows; the best probabilities for modeling sequence O are obtained by weighting the observations O_t using the residence probabilities $\gamma_t(i)$, for example, re-update $b_j(k)$ as

$$\hat{b}_j(k) = \frac{\sum_{i=1}^{T} \gamma_t(i) * (1 \text{ if } O_t = k, 0 \text{ otherwise})}{\sum_{i=1}^{T} \gamma_t(i)}$$

3. PROSODY AND EXPLICIT DURATIONS

In order to understand prosody, a unit called a phoneme must first be introduced. Phonemes are the basic sounds that make up the pronunciations of words in a language. A few examples of different phonemes are provided in figure 2. In ASR, phonemes are used to create a dictionary for the recognizer. An example dictionary entry is shown in figure 3(a).

Prosody is defined as any property of speech that is not limited to a specific phoneme. Two examples of prosody are accent and phrase boundary. We will not be dealing with accent in this paper and will concentrate mainly on the effects of phrase boundaries on different phonemes.

Tone and Break Indices (ToBI) is a labeling system used to mark accent and phrase boundaries. In the corpus used for our experiments, all word boundaries were labeled using ToBI. There are 5 different boundary labels (written as numbers 0-4) that can be given to a word. These labels determine what type of phrase boundary the word precedes. For example, an index of 1 indicates a phrase medial word and an index of 4 indicates the end of an intonational phrase.

As discussed in section 1, boundary phonemes can have a longer duration than non-boundary phonemes. Ferguson's algorithm was used to incorporate explicit duration densities into the HMM definitions. The exact algorithm used is discussed in detail in [4].

4. EXPERIMENTS

4.1 HTK

The HTK software was used for the purposes of training and testing our models. HTK is available from the University of Cambridge's Engineering Department [6]. HTK comes equipped with several tools for performing recognition experiments. The two tools we used for these experiments were HERest and HVite.

Each tool has a specific function. HEReset performs a single retraining of HMM parameters using the Baum-Welch algorithm. HVite is generally used as a Viterbi recognizer.

4.2 The Models

The experiments were conducted using data from speakers F1A and F2B from the Boston University Radio News Corpus. Five different acoustic models were created:

- Prosody Independent (IND)
- Final Syllable (FS)

- Final Rhyme. (FR)
- Initial Syllable. (IS)
- Initial Rhyme. (IR)

The name of each model describes which phonemes in a word are marked as being boundary phonemes. For example, IND has no phonemes marked as being boundary phonemes while FS marks all the phonemes in the final syllable of the word before a boundary as being boundary phonemes. Each different model had both a duration independent (DI) version and a duration dependent (DD) version that were trained for recognition testing.

The HMMs used contained three emitting states plus a non-emitting start and end state. Each emitting state included three Gaussian mixtures.

4.3 Transcriptions and Dictionaries

Radio News is one of the few speech databases made specifically for the study of prosody. Included in the corpus are files (file extension .brk) that contain ToBI labels for certain utterances in the corpus. We used these .brk files along with the Perl programming language and the included Radio News transcription files (file extension .wrd) to create prosody dependent word level transcriptions for use with HTK. Figure 4 shows an example of the .brk and .wrd files found in Radio News.

For our experiments, we had to make a simplification to the ToBI labeling due to the relatively small size of the Radio News Corpus. We considered any word that had an index of 3 or less to be a phrase medial word (have an index of 1) and we considered any word with an index of 4 to be a boundary word.

There were three types of word level transcriptions created. The first was the independent transcription file. This file contained no words marked as boundary words. The second transcription file marked only pre-boundary words as being affected by a prosodic boundary. The third transcription file marked only post-boundary words as prosodically affected words and all other words were marked as phrase medial. Prosodically affected words were marked with a "4." For example, "wrd" would be the phrase medial representation of some generic word whereas "wrd4" would be the phrase boundary representation of the same word. The actual pause for the boundary was represented in all three transcription files as BRK. Figure 5 shows examples of all our prosody independent and dependent transcriptions.

Phonemes were also separated into the categories of prosody independent and prosody dependent. In the prosody dependent models, each phoneme is modeled twice, once as a phrase medial phoneme and once as a boundary phoneme. We differentiated between boundary and non-boundary phoneme models by labeling boundary phonemes with a "4." (Figure 6)

Of our five models, the IND model was the only one to use the prosody independent transcriptions. FS and FR modeled phrase final phonemes so these models were trained and tested using the pre-boundary transcriptions. IS and IR modeled phrase initial phonemes so these models were trained and tested using the post boundary transcriptions.

In order to perform training and testing using our prosodic labeling of phonemes, we had to create special dictionaries for each of our four prosody dependent models. In the Radio News Corpus, pronunciations for words are given in the .prn files. We used a Perl script to read these files and output five distinct dictionaries, a prosody independent dictionary and four unique dictionaries with different prosodic labelings that corresponded to our four prosody dependent models.

Figures 3(b)-3(e) show examples of the dictionaries used by each of the four prosody dependent models.

The first experiment we ran was designed to find out whether the use of prosody as a "hidden" variable can help word recognition accuracy. In order to measure word recognition accuracy, the recognizer should output the prosody independent spelling each word every time it recognizes the word regardless of whether it thought the word was phrase-medial or phrase-final. The square brackets in figures 3(b)-3(e) indicate to HTK what information should be output when that particular word is recognized.

Recall that at the word level, models that were marked for pre-boundary prosodic effects shared the same transcriptions. Models marked for post-boundary effects also shared the same word level transcriptions. We used HVite in forced alignment mode to create phone level transcriptions for each of our models. Figure 6 shows examples of each of the five phone-level transcriptions used.

4.4 Word Recognition

The first experiments performed were standard word recognition experiments. The models described in section 4.2 were trained and tested using the transcriptions and dictionaries described in section 4.3.

Models from the DI set were re-estimated using HERest. Correctness and accuracy were tested for these models by using the tool HVite.

The HTK tools HERest and HVite had to be modified to incorporate explicit durations into the training and testing of the HMMs. The DD set was trained and tested using these modified versions of HERest and HVite.

4.5 Prosody Recognition

Prosody recognition experiments were run to determine whether or not the prosody dependent models were actually picking up on the prosodically marked phonemes. This test was run on the same models used for word recognition so no retraining was required.

The dictionary was modified for this test. Instead of having HTK output what word it had recognized, we had it output whether or not the recognized word had been prosodically marked. This allowed us to see whether the recognizer could distinguish between prosodically marked phonemes and words. Figure 7 shows examples of the modified dictionary and the output transcription. Any word marked for prosody was identified in the dictionary as a "B4" and any prosody independent word was identified ad "B0."

5. RESULTS

It can be seen clearly from table 1 that the use of prosody improves the correctness and accuracy of the standard HMM speech recognizer. Improvements due to prosody range from 2.53% to 3.99% in correctness and from 2.35% to 3.28% in accuracy. Explicit durations also help to improve a recognizer's correctness by as much as 1.32% and a recognizer's accuracy by as much as .85%. By comparing table 1 and table 2, we see that combining both prosody and explicit durations yields a much greater improvement in both correctness and accuracy than simply incorporating either prosody or explicit durations in the model.

Our experiments using prosody have also revealed that there is information a recognizer can make use of in the entire pre or post boundary syllable. While all the prosody based models outperformed the prosody independent model, the models which assume there is prosodic information in the entire syllable out-perform the models that only assume there is prosodic information in the rhyme of the syllable.

The results in tables 3 and 4 confirm that the use of prosody allows a recognizer to pick out words on a phrase boundary. The recognizer was able to pick out 75% to 83% of the boundary words depending on the model. This indicates that there is a usable difference between regular phonemes and boundary phonemes. A comparison of tables 3 and 4 shows that explicit durations help the recognizer to pick up on prosodic information.

Results from tables 3 and 4 also seem to indicate that there is more usable prosodic information in the final phonemes of pre-boundary words than there is in the initial phonemes of post-boundary words. For example, IS performs 3.49% lower in correctness and 3.49% lower in accuracy than the FS model. Similarly, the duration independent IR model performs lower than FS in both correctness and accuracy for prosody recognition as well.

6. CONCLUSION

In this paper, we experimented with the use of prosody and explicit durations in the standard HMM speech recognizer. For the models we created, we ran experiments to check word recognition correctness and accuracy. We also ran experiments to confirm whether or not recognizer could pick up on the prosody of boundary syllables.

We have shown two effective methods for improving both the correctness and accuracy of an HMM speech recognizer. The first method involves using prosody to distinguish certain phonemes in pre or post boundary words. The second method involves modifying the HMM definition so that it can use explicit durations. Combining these two methods can further improve recognition rates.

We have also shown that the entire syllable of a pre or post boundary word is affected by a boundary. While post boundary information appears to be useful, it does not appear to be as useful as pre-boundary information.

7. ACKNOWLEDGEMENTS

This work was supported by a grant from the University of Illinois Critical Research Initiative.

8. REFERENCES

[1] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp 1707-1717, March 1992.

[2] C. Fougeron and P. A. Keating, “Articulatory strengthening at edges of prosodic domains,” *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3728-3740, June 1997.

[3] Kompe, “Prosody in speech understanding systems,” *Lecture Notes in Artificial Intelligence,* vol. 1307, pp. 1-357, July 1997.

[4] K. Chen, S. Borys, and M. Hasegawa-Johnson, “A prosody dependent speech recognizer for the study of lengthening in the vicinity of intonational phrase boundary,” (in preparation)

[5] L. R. Rabiner and B. H. Jaung, “An introduction to hidden Markov models,” *IEEE ASSP Magazine*, vol. 1, pp. 4-15, January 1986.

[6] The University of Cambridge Engineering Department, “http://htk.eng.cam.ac.uk/”, December 2002

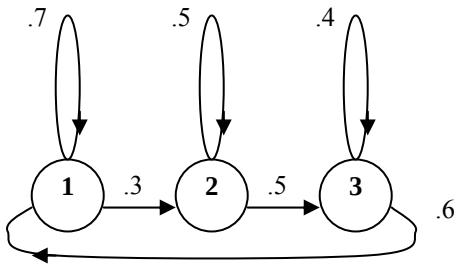

Figure 1. A simplified example of a hidden Markov model.

ae	b<u>a</u>t
ax	th<u>e</u>
b	<u>b</u>ob
ow	b<u>oa</u>t
th	<u>th</u>ief

Figure 2. Five examples of different phonemes. In the standard phoneme set, there are 48 different phonemes

(a) `center s eh n cl t ax r`

(b) `center4 [center] s eh n cl t4 ax4 r4`

(c) `center4 [center] s4 eh4 n4 cl t ax r`

(d) `center4 [center] s eh n cl t ax4 r4`

(e) `center4 [center] s eh4 n4 cl t ax r`

Figure 3. (a) An example from the prosody independent dictionary. **(b)** An example from the FS dictionary. **(c)** An example from the IS dictionary. **(d)** An example from the FR dictionary. **(e)** An example from the IR dictionary.

(a)

```
23650000 23820000 their
23820000 24150000 own
24150000 24150000 BRK
24150000 24710001 what
24710001 24790001 do
```

(b)

```
23650000 23820000 their
23820000 24150000 own4
24150000 24150000 BRK
24150000 24710001 what
24710001 24790001 do
```

(c)

```
23650000 23820000 their
23820000 24150000 own
24150000 24150000 BRK
24150000 24710001 what4
24710001 24790001 do
```

Figure 5. (a) A sample word level transcription for a prosody independent model. **(b)** A sample pre-boundary transcription. **(c)** A sample post-boundary transcription.

(a)	**(b)**	**(c)**	**(d)**	**(e)**
t	t	t	t	t
uw	uw	uw	uw	uw
vcl	vcl	vcl	vcl	vcl
d	d4	d	d	d
ey	ey4	ey	ey4	ey
sil	sil	sil	sil	sil
vcl	vcl	vcl	vcl	vcl
b	b	b4	b	b
ae	ae	ae4	ae	ae4
r	r	r4	r	r4
iy	iy	iy	iy	iy

Figure 6. (a) A sample phone level transcription for a prosody independent model. **(b)** A sample phone level transcription from the FS model. **(c)** A phone level transcription from the IS model. **(d)** A sample phone level transcription from the FR model. **(e)** A sample phone level transcription from the IR model. The words transcribed are "today BRK barry." BRK represents the word boundary and is represented with silence.

(a)

```
29.120001 76 one
29.240000 76 in
29.309999 76 a
29.850000 76 series
30.020000 76 of
30.360001 76 recent
31.139999 76 anti-tax
```

(b)

```
29.120001 76 1
29.240000 76 0
29.309999 76 1
29.850000 76 4
30.020000 76 1
30.360001 76 2
31.139999 76 4
```

Figure 4. (a) An example of a .wrd file from Radio News. The first column is the time in seconds when the word ends. The third column is the word. **(b)** An example of a .brk file from Radio News. The column three numbers are the ToBI break indices.

Duration Independent Word Recognition		
Model	**%Corr**	**%Acc**
IND	49.32	46.44
FS	52.35	49.18
FR	51.85	48.79
IS	53.31	49.72
IR	53.24	48.93

Table 1 Results from duration independent word recognition.

Duration Dependent Word Recognition		
Model	**%Corr**	**%Acc**
IND	49.86	46.90
FS	53.38	49.96
FR	53.17	49.64
IS	54.31	50.11
IR	53.84	49.54

Table 2 Results from duration dependent word recognition.

Duration Independent Prosody Recognition		
Model	**%Corr**	**%Acc**
IND	-	-
FS	81.57	80.00
FR	80.00	78.26
IS	78.08	76.51
IR	75.59	73.74

Table 3 Results from duration independent prosody recognition.

Duration Dependent Prosody Recognition		
Model	**%Corr**	**%Acc**
IND	-	-
FS	83.13	81.39
FR	81.60	79.93
IS	80.71	78.90
IR	78.90	76.62

Table 4 Results from duration dependent prosody recognition.

(a)

```
center [B0] s eh n cl t ax r
center4 [B4] s eh n cl t4 ax4 r4
```

(b)

```
23650000 23820000 B0
23820000 24150000 B4
24150000 24150000 B0
24150000 24710001 B0
```

Figure 7. (a) An example of the modified dictionary used for prosody recognition. **(b)** An example of output generated by HVite during the prosody recognition tests.