Modeling Prosodic Consistency for Automatic Speech Recognition: Preliminary Investigations

Ernest Pusateri and James Glass
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{pusateri, glass}@mit.edu

Abstract

In this paper we describe a prosody-dependent duration model as a first step toward incorporating a prosodic consistency constraint into a speech recognizer. As part of this model, we describe a text-based prosody prediction scheme, novel in its use of a preliminary integrated comma-prediction/POS tagging step. We also demonstrate a relative decrease in perplexity using the prosody-dependent duration model and analyze what conditioning factors most contributed to that decrease. The analysis indicates that while word position is, by far, the most important factor, predicted prosodic labeling information also contributes to the decrease. This final result suggests a benefit to integrating a prosodic consistency constraint into a speech recognition system.

1 Introduction

While much effort has gone into using prosody in the areas of speech synthesis and understanding (e.g (Noth et al., 2000; Taylor and Black, 1998)), less has been focused on using it to aid directly in the task of speech recognition (e.g. (Stolcke et al., 1999; Ostendorf et al., 2003; Chen et al., 2003).) The utility of prosody in speech recognition comes from the fact that, while prosody is not fully determined by an utterances lexical content, lexical content does make some prosodic realizations more probable than others. This implies that we can meaningfully ask whether the acoustic cues to the prosody of an utterance are consistent with a textual hypothesis.

In this work we report on an initial effort to develop a prosody-dependent duration model. It is closely related to (Chen et al., 2003). However, while that work uses a standard language model for text-based prosody prediction, we incorporate techniques borrowed from speech synthesis research as well as an automatic comma annotation technique in an effort to increase robustness.

The rest of this paper proceeds as follows. In Section 2 we begin with a description of a very general framework to incorporate a prosodic consistency constraint into speech recognition. In Section 3, we describe our text-based prosody prediction algorithm. This is followed by a description of duration modeling in Section 4. After that, we present experiments and results in Section 5. We end with a brief summary in Section 6 and a discussion of future work in Section 7.

2 A Prosodic Consistency Framework

Figure 1 illustrates a framework for incorporating a prosodic consistency constraint into a speech recognizer. On the left path, an N-best list is generated by the recognizer and text-based prosody prediction is performed on each of the N-best entries. On the right path the utterance is analyzed for acoustic-prosodic cues. The level of consistency between the predicted prosody for each entry and the acoustic cues measured in the utterance is then used to rescore the N-best list.

This work does not implement a speech recognizer. Instead, our intent is to show that using a particular prosodic cue in this framework, specifically duration, has the potential to reduce the recognition search space. In this effort, the "Acoustic Prosodic Analysis" component shown in Figure 1 simply reads the phone durations from the N-best list. The "Text-based Prosody Prediction" component, however, is fairly complex, and the next section gives a full description of its inner-workings.

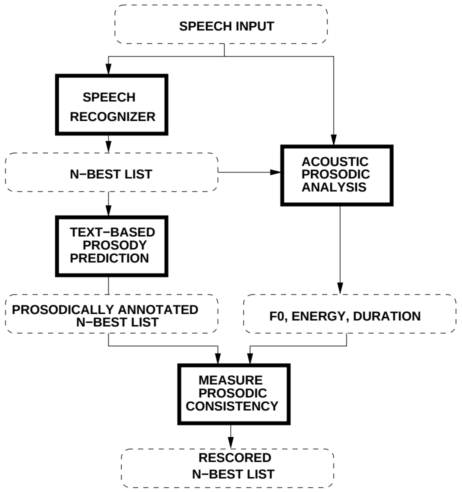
SPEECH INPUT
SPEECH RECOGNIZER
N-BEST LIST
ACOUSTIC PROSODIC ANALYSIS
TEXT-BASED PROSODY PREDICTION
PROSODICALLY ANNOTATED N-BEST LIST
F0, ENERGY, DURATION
MEASURE PROSODIC CONSISTENCY
RESCORED N-BEST LIST

Figure 1: A framework for incorporating prosodic consistency into a speech recognition system.

3 Text-based Prosody Prediction

3.1 Overview

The problem being addressed by text-based prosody prediction is that of assigning probabilities to prosodic labellings for a string of text, that is we are attempting to find the probability, $P(L|W)$ for a particular prosodic labeling, L, given the word string, $W = \{w_1, w_2 ... w_N\}$. The labeling consists of phrase breaks and prominences. Specifically, each word boundary is put into one of three categories: no phrase break, intermediate phrase break, or full phrase break. Each word is labeled as either prominent or not prominent.

We decompose the probability $P(L|W)$ based on comma locations within the text, C, and part-of-speech (POS) labels, S:

$$\begin{aligned} P(L|W) &= \sum_{C,S} P(L,C,S|W) && (1) \\ &= \sum_{C,S} P(L|C,S,W)P(C,S|W) && (2) \end{aligned}$$

This decomposition suggests a comma and POS prediction step to obtain $P(C,S|W)$ followed by prosodic prediction to obtain $P(L|C,S,W)$. The method is novel in the inclusion of the comma prediction step. The motivation for this comes from the observation that full phrase breaks often occur in the presence of a comma. Indeed, predicting prosodic boundaries with punctuated text is an easier task (Ostendorf and Veilleux, 1994; Taylor and Black, 1998). An alternative approach, taken in (Ostendorf and Veilleux, 1994), would be to do syntactic analysis of the text, and then use constituent boundaries to aid in boundary prediction. However, we hope that our method may later be successfully applied to more spontaneous speech, which is difficult to parse syntactically. We will now describe the models used for $P(C,S|W)$ and $P(L|C,S,W)$.

3.2 Comma Prediction and POS Labeling

With the comma prediction model, we are attempting to compute the probability $P(C,S|W)$, that is the probability of a particular set of comma placements, C, and POS tags, S, for a word string, W. By jointly determining C and S, this implementation represents a hybrid between the comma prediction algorithm presented in (Beeferman et al., 1998) and an HMM POS tagging algorithm like that in (DeRose, 1988).

The hybrid model is separated into a unigram POS component capturing $P(W|S)$, the probability that a particular POS tag is attached to a particular word, and a comma/POS N-gram component representing $P(C,S)$:

$$\begin{aligned} P(C,S|W) &= \frac{P(C,S,W)}{P(W)} && (3) \\ &= \frac{P(W|C,S)P(C,S)}{P(W)} && (4) \\ P(W|C,S) &\approx P(W|S) && (5) \\ &= P(w_1^N|s_1^N) && (6) \\ &\approx \prod_{i=1}^{N} P(w_i|s_i) && (7) \end{aligned}$$

The history for the $p(C,S)$ N-gram is based on the past N-1 POS labels and/or commas.

Underlying this implementation is the assumption that comma locations are independent of the words given the words' parts-of-speech. This is motivated by the fact that commas occur at syntactic breaks, and syntax and part-of-speech are very strongly related. Still, this represents a very large assumption. In Section 5.3.1 we discuss how this assumption is relaxed by giving common words their own POS classes.

3.3 Prosodic Prediction

In the prosodic prediction component of the system, we are computing the probability $P(L|C,S,W)$. We separate L into two parts: break classifications (B) and prominence classifications (R).

$$\begin{aligned} P(L|C,S,W) = &\quad P(B,R|C,S,W) && (8) \\ = &\quad P(R|B,C,S,W)P(B|C,S,W) && (9) \end{aligned}$$

We will first discuss how we obtain the break term, $P(B|C,S,W)$, then how we obtain the prominence term, $P(R|B,C,S,W)$.

3.3.1 Break Prediction

In obtaining $P(B|C,S,W)$, we first assume that B is conditionally independent of W given S and C:

$$P(B|C,S,W) \approx P(B|C,S) \tag{10}$$

While this assumption results in ignoring word-specific (and thus also semantic) cues to break location, it does allow us to use some syntactic information. Both (Ostendorf and Veilleux, 1994) and (Taylor and Black, 1998) have found this approximation to be workable.

We relate the approximation to the joint probability, $P(B,C,S)$:

$$P(B|C,S) = \frac{P(B,C,S)}{P(C,S)} \tag{11}$$

To find $P(B,C,S)$, we can use the framework presented in (Taylor and Black, 1998). Let b_i be the type of the boundary between w_i and w_{i+1}.

$$\begin{aligned} P(B,C,S) &= P(C,S|B)P(B) & (12)\\ &= (\prod_{i=1}^{N} P(c_i,s_i,s_{i+1}|B))P(B) & (13)\\ &\approx (\prod_{i=1}^{N} P(c_i,s_i,s_{i+1}|b_i))P(B) & (14)\end{aligned}$$

The first component of the model, $P(c_i,s_i,s_{i+1}|b_i)$ captures the distribution of the parts-of speech surrounding boundaries. The second component, $P(B)$, captures common boundary type patterns and is modeled with an N-gram. The normalization term in Equation 11, $P(C,S)$ is estimated from the training data.

The parts of speech were collapsed into the classes described in (Ostendorf and Veilleux, 1994): content words, determiners, prepositions, and a general class incorporating function words that were neither determiners nor prepositions. This is a smaller set than was used in (Taylor and Black, 1998), and seemed more appropriate given that we were working with a smaller amount of training data.

3.3.2 Prominence Prediction

A simple model was used to compute $P(R|B,C,S,W)$. The word classes used for break prediction were further collapsed into function and content word classes. A unigram model of prominence based on these classes was then applied. Thus, the following assumption was made:

$$\begin{aligned} P(R|B,C,S,W) &\approx P(R|S) & (15)\\ & & (16)\end{aligned}$$

This assumption is motivated by the fact that content words very often are prominent while function words very often are not. While this simple unigram model tends to over-predict prominences, it is used by many speech synthesizers.

4 Duration Modeling

In this work we model only vowel durations. The end result of applying the duration model should be the probability of the vowel durations, $D = \{d_{1,1}, d_{1,2}..., d_{2,1}, d_{2,2}, ...\}$ (where $d_{i,j}$ corresponds to the jth vowel in w_i), given the word sequence, W.

To allow the incorporation of prosodic prediction, we decompose $P(D|W)$:

$$\begin{aligned} P(D|W) &= \sum_{P} P(D,L|W) & (17)\\ &= \sum_{L} P(D|W,L)P(L|W) & (18)\end{aligned}$$

$P(L|W)$ is the probability computed by our text-based prosody prediction model.

The assumption is made that the duration $d_{i,j}$ depends only on w_i (the word to which the vowel belongs), b_i (the type of the following boundary) and p_i (whether or not word i is prominent). Also, we assume that, given the word string and prosodic labeling, the durations are independent. This gives us:

$$P(D|W,L) \approx \prod_{i=1}^{N} \prod_{j=1}^{M_i} P(d_{i,j}|w_i,b_i,l_i) \tag{19}$$

Raw durations are normalized for both speaking rate and vowel identity, using the method described in (Wightman et al., 1992). This normalization makes the duration independence assumption reasonable. Normalized durations are modeled as Gaussian distributions, and separate models are built depending on 4 factors. The first factor is the *lexical stress* of the vowel. Second is the vowel's *word position* (i.e. whether or not the vowel is in the last syllable of the word.) These first two factors reflect the dependence of duration on w_i. The third factor is the *boundary type* following the word (i.e. whether the word precedes an intermediate phrase break, a full phrase break, or no phrase break.) This reflects the dependence of duration on b_i. The final factor is *prominence* (i.e. whether or not the word containing the vowel is prominent.) This factor reflects the dependence of duration on l_i.

5 Experiments and Results

5.1 Data

Training and testing of the comma/POS prediction component were completed using the tagged Wall Street Journal portion of the Treebank corpus (Marcus et al., 1993). 130,226 utterances were used for training, while 1,986 were used for testing.

Training and testing of the prosodic prediction component as well as the duration model were completed using the FM Radio News corpus (Ostendorf and Veilleux, 1994). 485 (3 news stories read by 5 speakers) utterances were used for training, a superset of the 312 used in (Ostendorf and Veilleux, 1994). For prosodic prediction, 23 sentences were used for testing with 5 possible prosodic transcriptions considered correct. This is the same test set used in (Ostendorf and Veilleux, 1994). The same 23 sentences read by a single speaker were used for duration model tests. While the test speaker was part of the training data, the test news story was not.

5.2 Evaluation Metric

In order to evaluate the extent to which duration modeling was constraining the recognition search space, we derived a measure of perplexity reduction. In its standard form, perplexity measures the uncertainty present in a language model. We wanted a measure of how much prosody-dependent duration information reduced uncertainty.

Suppose we computed perplexity using $P(W|D)$ instead of $P(W)$:

$$PP_{dur} = 2^{\frac{-1}{N} log_2 P(W|D)} \tag{20}$$

To obtain $P(W|D)$, we can use Bayes rule with the probability computed by the duration model (see Section 4):

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)} \tag{21}$$

If we wish to compare the results of two duration models, a and b, we can look at the percentage by which model b reduces this duration-dependent perplexity:

$$\begin{aligned} R_{PP} &= 1 - \frac{PP^{(b)}_{dur}}{PP^{(a)}_{dur}} && (22) \\ &= 1 - \frac{2^{\frac{-1}{N} log_2 P^{(b)}(W|D)}}{2^{\frac{-1}{N} log_2 P^{(a)}(W|D)}} && (23) \\ &= 1 - \left(\frac{P^{(a)}(D|W)}{P^{(b)}(D|W)}\right)^{\frac{1}{N}} && (24) \end{aligned}$$

Thus our evaluation metric, R_{PP}, can be computed directly from our duration model probability and a baseline duration model probability. For the baseline, we use a global model trained on all vowel data without regard to lexical stress, word position, or phrase break or prominence locations, although speaking rate and vowel identity normalization were still performed.

5.3 Results

5.3.1 Text-based prediction

We first evaluate the performance of the comma prediction component of the system. A 61.3% recall rate and a 3.0% false dection rate are obtained, where the recall rate is the probability that a comma is predicted at a word boundary with a comma, and the false detection rate is the probability that a comma is predicted where none exists in the transcription. In this experiment, a 5-gram model was utilized and the top 1000 words/POS pairs (accounting for 51.2% of the words in the training data) were assigned special POS tags.

Now we turn to phrase break prediction results, shown in Table 1. As mentioned previously, 5 "correct" prosodic labellings were available for each of the test utterances, corresponding to the realizations of 5 different speakers. The labeling most similar to the automatic labeling for each utterance was used to compute the results in the table. About 7% of boundaries were full phrase breaks. For computational reasons, Equation 2 was not implemented as is. Instead of summing over all C, the highest probability comma annotation was chosen and used in the prosodic prediction step.

The table shows results under three different conditions: using transcribed commas, without using any comma prediction and using predicted comma locations (from the model using POS information.) We see that, while the system using predicted comma locations does not perform as well as the one using the transcribed comma locations, it does perform better overall than the system without comma prediction.

We can also compare these results to those reported in (Ostendorf and Veilleux, 1994), labeled O & V in the table. We see that, using transcribed comma information, our system, which, under this condition, is virtually identical to (Taylor and Black, 1998), achieves a higher recall rate, but at the cost of a higher false detection rate. Similarly, our system using predicted commas has a higher recall rate than the first O & V system, but it also has a higher false detection rate. Finally, considering that only about 7% of the boundaries are phrase breaks, it does not

Model	Commas	R	FD
-	None	69.4	8.3
-	Transcribed	87.0	6.0
-	Predicted	75.4	7.6
O & V	None	66	5
O & V (w/syntax)	None	71	4
O & V	Transcribed	81	4

Table 1: Full phrase break prediction recall and false detection percentages without comma information, with transcribed comma information and with predicted comma information. Results labeled O & V are taken from (Ostendorf and Veilleux, 1994).

appear to perform as well overall as the O & V system that incorporates syntax. This was expected, as, in the O & V system the syntax is hand transcribed.

Our simple unigram model for prominence prediction achieved 78.7% recall and 41% false detection.

5.3.2 Duration Modeling

Now we use the evaluation metric described in Section 5.2 to assess whether or not we can use these prosodic differences in duration to aid in recognition. The results are shown in Table 2. Values for R_{PP} are given both using the labeled prosody of the test data as well as the prosodic labeling predicted by the text-based model. The first row contains values of R_{PP} computed using all of the duration conditioning factors enumerated in Section 4. The value of the metric suggests a significant decrease in uncertainty.

The values of R_{PP} in the remaining rows are computed by removing one conditioning factor. This gives us an idea of how much each factor contributed to the value in the first row. We see that, by far, word position is the most important conditioning factor. Removing it results in a sharp decrease in R_{PP}. Information about phrase break location has the next most significant effect, with its removal resulting in decreases of .03 and .02 in the value of R_{PP} in the labeled and predicted cases respectively. Prominence is next, showing decreases of .03 and .01, while removing lexical stress as a conditioning factor results in a decrease of only .01 in both cases.

We were somewhat surprised that word position was so important in comparison to break location. We see two possible reasons for this. First, word position affects one vowel in every word in every test utterance. Phrase breaks occur only after about one fifth of the words, resulting in less impact on the probability $P(D|W)$. Second, the speech in the corpus was read by professional radio announcers, whose job involves being exceptionally intelligible. We speculate that this may make durational differences less drastic than they may be in more casual speech.

	R_{PP}	
Conditioning Factors	Labeled Prosody	Predicted Prosody
all	.16	.14
-word position	.03	.03
-break	.13	.12
-prominence	.13	.13
-lexical stress	.15	.13

Table 2: Reduction in conditional perplexity using a vowel duration model conditioned on word position, break, prominence, and lexical stress. Dependence on factors is removed one at a time to gauge the importance of each.

Table 2 also shows a decrease in R_{PP} when we move from using labeled prosody to using predicted prosody. This was expected. Even the best text-based prosody model could not predict the exact prosodic realization of a particular text string, as it is an inherently ambiguous task. That said, our text-based model could certainly be improved. Still, the predicted prosody-dependent factors show some effect on R_{PP}.

6 Summary

In this work we have implemented a text-based prosody prediction scheme, novel in its use of a preliminary integrated comma-prediction/POS tagging step. We have also demonstrated an increase in a word-level constraint metric using prosody-dependent duration models, and analyzed what conditioning factors most contributed to that increase. The analysis indicates that while word position is, by far, the most important factor, predicted prosodic labeling information also contributes to the increase. This final result suggests a benefit to integrating prosodic consistency into a speech recognition system.

7 Future Work

The ultimate goal of this work is to use prosodic consistency as a constraint in a speech recognizer. To this end, we plan to close the loop on the work described here by incorporating prosody-based duration modeling into a speech recognition system. We also plan to incorporate more acoustic prosodic cues including pause duration, and fundamental frequency information into this framework.

We feel that prosodic consistency may provide an especially valuable constraint in more casual speech. With this in mind, we are looking to move away from the read speech domain used here and into more spontaneous domains like university course lectures and, at the extreme, phone conversations.

References

D. Beeferman, A. Berger, and J. Lafferty. 1998. CYBERPUNC: A lightweight puncutation annotation system for speech. In *Proc. ICASSP*, pages 689–693, Seattle.

K. Chen, S. Borys, M. Hasegawa-Johnson, and J. Cole. 2003. Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. In *Proc. Eurospeech*, pages 393–396, Geneva.

S. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14:31–39.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(1).

E. Noth, A. Batliner, A. Kiebling, R. Kompe, and H. Niemann. 2000. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio Processing*, 27:113–134.

M. Ostendorf and N. Veilleux. 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20:27–54.

M. Ostendorf, I. Shafran, and R. Bates. 2003. Prosody models for conversational speech recognition. In *Proc. for 4th Plenary Meeting and Symposium on Prosody and Speech Processing.*

A. Stolcke, E. Shriberg, D. Hakkani-Tur, and G. Tur. 1999. Modeling the prosody of hidden events for improved word recognition. In *Proc. Eurospeech*, pages 311–314.

P. Taylor and A. Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117.

C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am.*, 91(3):1707–1717.