Exploiting Prosodic Breaks in Language Modeling with Random Forests

Yi Su and Frederick Jelinek

Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University, Baltimore, Maryland, USA
{suy; jelinek}@jhu.edu

Abstract

We propose a novel method of exploiting prosodic breaks in language modeling for automatic speech recognition (ASR) based on the random forest language model (RFLM), which is a collection of randomized decision tree language models and can potentially ask any questions about the history in order to predict the future. We demonstrate how questions about prosodic breaks can be easily incorporated into the RFLM and present two language models which treat prosodic breaks as observable and hidden variables, respectively. Meanwhile, we show empirically that a finer grained prosodic break is needed for language modeling. Experimental results showed that given prosodic breaks, we were able to reduce the LM perplexity by a significant margin, suggesting a prosodic N-best rescoring approach for ASR.

1. Introduction

Prosody refers to a wide range of suprasegmental properties of spoken language units, including tone, intonation, rhythm, stress and so on. It has been used for a number of spoken language processing tasks, such as disfluency and sentence boundary detection [1], topic segmentation [2], spoken language parsing [3], among others. We are mainly interested in using prosody to improve automatic speech recognition (ASR). As a separate knowledge source, prosody has been helpful in all three major components of a modern ASR system: the acoustic model [4, 5], the pronunciation model [6] and the language model [7, 8]. (For a comprehensive review of prosody models in ASR, see [9].) New opportunities of using prosody emerged after the availability of a prosodically labeled speech corpus [10], where tones and breaks were hand-labeled with a subset of ToBI labeling scheme [11]. In this work, we focus on prosodic breaks.

The random forest language model (RFLM) [12] is a powerful model which consistently outperforms the n-gram language model in terms of both perplexity and word error rate in several state-of-the-art ASR systems [13, 14]. Based on decision trees, the RFLM has the potential of integrating information from various sources besides the history words by simply asking new questions, analogous to the maximum entropy language model by using new features [15, 16]. We propose two prosodic language models based on the RFLM and demonstrate their performance in perplexity by contrasting them to a baseline n-gram language model using the same information.

The rest of the paper is organized as follows: in Section 2 we present our proposed models. In Section 3, we briefly review the RFLM. Experimental setup and results are presented in Section 4. Discussion of future work appears in Section 5 and conclusions in Section 6.

2. Prosodic Language Models

2.1. Granularity of Prosodic Breaks

The ToBI-labeled speech corpus [10] makes it possible to investigate the use of prosodic breaks in two aspects: automatic detection/classification and statistical modeling. Although some researchers argued against this intermediate phonological layer [17], we believe that 1) supervised training of prosodic classifiers can help us understand the usefulness of various proposed prosodic features; 2) symbolic prosodic breaks are easier to fit into the current n-gram based language modeling approach than continuous-valued prosodic features. As an example of the second point, in direct modeling of prosodic features like the pause length in [18], simple quantization like binning was used.

A decision tree classifier was built to predict three types of breaks, namely, 1, 4 and p, with an accuracy of 83.12% in [3] for parsing speech. In [19], the 1 and p labels were further collapsed into one. While this granularity of prosodic breaks has been suitable for their tasks, we believe a finer granularity is needed for language modeling. So we used the quantized posterior probability $P(1|\text{features})$, which has 12 possible values, from the decision tree classifier of [3] in our experiments. (See Section 4.2 for details.)

2.2. Language Models with Prosodic Breaks

Let $W, S = w_0 s_0 w_1 s_1 w_2 s_2 \cdots w_m s_m$ be a sequence of words and prosodic breaks in their temporal order, where $W = w_0 w_1 w_2 \cdots w_m$ is the sentence of length $(m+1)$ and $S = s_0 s_1 s_0 \cdots s_m$ is the sequence of prosodic breaks for the sentence W, where s_i denotes the break between w_i and w_{i+1}, for all $0 \leq i < m$.

First we would like to estimate the joint probability $P(W, S)$ as an n-gram LM of the $(\text{word}, \text{break})$-tuples:

$$\begin{aligned} P(W,S) &= \prod_{i=0}^{m} P(w_i, s_i | w_0^{i-1}, s_0^{i-1}) \\ &\approx \prod_{i=0}^{m} P(w_i, s_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}). \end{aligned} \quad (1)$$

This joint model is immediately usable if our goal of ASR is the simultaneous recognition of words and prosodic breaks [20, 5]:

$$\begin{aligned} (W,S)^* &= \arg\max_{W,S} P(W,S|A) \\ &= \arg\max_{W,S} P(A|W,S)P(W,S) \\ &\approx \arg\max_{W,S} P(A|W)P(W,S), \end{aligned} \quad (2)$$

where A stands for the acoustic features and we approximate $P(A|W,S)$ with a usual acoustic model $P(A|W)$ for the sake of simplicity[1].

If we stick to the original formulation of ASR, we can estimate the language model $P(W)$ as follows:

$$\begin{aligned} P(W) &= \sum_S P(W,S) \\ &= \sum_S \prod_{i=0}^{m} P(w_i, s_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}). \end{aligned} \quad (3)$$

This computation can be carried out efficiently by a simple forward pass of the forward-backward algorithm [21].

For either (1) or (3), we need to compute the probability $P(w_i, s_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$. We propose the following two methods:

- Let $t_i = (w_i, s_i)$, for all $0 \le i \le m$. We have

$$P(w_i, s_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}) = P(t_i | t_{i-n+1}^{i-1}). \quad (4)$$

  Then we can build an n-gram LM or RFLM of t_i's, whose vocabulary is the Cartesian product of the word vocabulary and the prosodic break vocabulary.

- Alternatively, we can decompose the probability as follows:

$$\begin{aligned} & P(w_i, s_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}) \\ = \; & P(w_i | w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}) \\ & \cdot P(s_i | w_{i-n+1}^{i}, s_{i-n+1}^{i-1}). \end{aligned} \quad (5)$$

  Then we can build two n-gram LMs or RFLMs for predicting the word and the break, respectively.

In the second method, however, when a history consists of both words and prosodic breaks, there isn't a natural order of backing off. Previous work either chose it heuristically (e.g., [22, 23]) or tried to find the optimal back-off path or combination of paths [24, 25]. We propose to handle this problem gracefully with the RFLM, which we will describe in the following section.

3. Random Forest Language Models

A RFLM [12] is a collection of randomized decision tree language models (DTLMs) [26], which define equivalence classification of histories. The RFLM generalizes the DTLM by averaging multiple DTLMs, which, in turn, generalizes the n-gram LM by having a sophisticated equivalence classification. The LM probability of a RFLM is defined as follows:

$$\begin{aligned} P_{RF}(w|h) &= \frac{1}{M} \sum_{j=1}^{M} P_{DT_j}(w|h) \\ &= \frac{1}{M} \sum_{j=1}^{M} P(w | \Phi_{DT_j}(h)), \end{aligned} \quad (6)$$

where h is the history and $\Phi_{DT_j}(\cdot)$ is a decision tree. The questions that have been used so far care only about the identity of the words in a history position. If w_i is the word we want to predict, then the question takes the following form:

> Is the word $w_{i-k}, k \ge 1$ in a set of words $\mathcal{S}$?

Because in the normal n-gram situation we know no more than the words in the history, these questions are almost all we can ask.

Now if we have more information about the history, we can easily enlarge our inventory to include questions of the following form:

> Does the feature f about the history take its value in a set of values $\mathcal{S}$?

As long as the feature values are categorical, we can use the same decision tree building algorithm as before. This makes the RFLM an ideal model framework for integrating information from various sources. For example, if we are given the prosodic breaks between words in the history, we can ask questions like:

> Does the prosodic break s_{i-1} take its value in the set of values $\{0.7, 0.8, 0.9\}$?

Note that from the decision tree's point of view, w_{i-k} is just another feature which happens to take its value in the vocabulary. Only when it is informative for the prediction do we want to ask questions about it. As numerous LMs, like the n-gram LM or the maximum entropy LM, have proven, the immediately previous word w_{i-1} is the single most important/informative and the most easily obtainable feature, followed by, probably, the previous of previous word w_{i-2}.

4. Experiments

4.1. Data and Setup

We used the ToBI-labeled Switchboard data from [10]. Following [3], we divided our data into training ($665,790$ words), development ($51,326$ words) and evaluation ($49,494$ words, $55,529$ counting the end of sentence symbols). Due to the relatively small size of the corpus, our LMs would only consider up to two words and two breaks in the history, if not specified otherwise. We built 100 trees for each RFLM and the smoothing method for both regular n-gram LMs and RFLMs was always the modified Kneser-Ney [27]. The vocabulary size was 10k.

4.2. Granularity of Prosodic Breaks

The decision tree classifier in [3] provided three degrees of granularity: two-level (break or not), three-level (ToBI indices 1, 4 and p) and continuous-valued (quantized into 12 values, $0.0, 0.1, \ldots, 1.0$ and -1.0). We built three RFLMs for $P(w_i | w_{i-1}, w_{i-2}, s_{i-1}, s_{i-2})$, where the breaks s_{i-1} and s_{i-2} took values of different granularity. The baseline was the word trigram LM, $P(w_i | w_{i-1}, w_{i-2})$, with modified Kneser-Ney smoothing.

Table 1: *Granularity of Prosodic Breaks*

Model	two-level	three-level	cont.-valued
KN.3gm	66.1	66.1	66.1
RF-100	65.5	65.4	**56.2**

From Table 1, we concluded that the ToBI indices were not fine-grained enough for the purpose of language modeling. Henceforth our experiments used the continuous-valued breaks.

[1] Another way to justify this approximation is that in this paper, we only consider breaks, among many other prosodic features, and prosodic breaks have a relatively weak influence on the acoustics.

4.3. Feature Selection by RFLM

As we mentioned before, from a RFLM's point of view, the various variables in the history, w_i's or s_i's, are just features. The model chooses any one of them simply because it has strong correlation, or large mutual information, with the future word. So by asking the RFLM *not* to use one of the variables in the history, we can find out how valuable that feature is. This kind of feature engineering was also used in maximum entropy LMs [15, 16].

We built RFLMs for $P(w_i|w_{i-1}, w_{i-2}, s_{i-1}, s_{i-2})$ then masked out one of the features in order to see how much it contributed.

Table 2: *Feature Selection by RFLM*

History	Perplexity
w_{i-1}, w_{i-2}, s_{i-1}, s_{i-2}	56.2
w_{i-1}, w_{i-2}, s_{i-1}	**55.9**
w_{i-1}, w_{i-2}, s_{i-2}	63.9
w_{i-1}, w_{i-2}	62.3

As we had expected, Table 2 showed that the break between the immediately previous word and the future word, s_{i-1}, helped the prediction, while the break between the previous and the previous of the previous, s_{i-2}, did not. Adding the latter actually hurt the perplexity a little bit, although that might change if we had more data. Similar experiments can be done for $P(s_i|w_i, w_{i-1}, w_{i-2}, s_{i-1}, s_{i-2})$. We skipped the detail but the conclusion was that the most useful features for predicting a break were its previous two words, w_i and w_{i-1}, which was consistent with our intuition.

We also point out here that this kind of experiments would *not* have been so easy to carry out in the case of regular n-gram LMs with modified Kneser-Ney smoothing. You have to specify the back-off order and search the best value for some of the discount parameters.

4.4. Main Perplexity Results

Having selected the features, we put the two components together following (5) to get $P(w_i, s_i|w_{i-1}, w_{i-2}, s_{i-1}, s_{i-2})$ and called it the "decomp." (decomposition) method in Table 3. For comparison, we also followed (4) to get the same quantity with a trigram LM of $(\text{word}, \text{break})$-tuples and called it the "tuple 3gm" method in Table 3. For each method, we contrasted the modified Kneser-Ney-smoothed n-gram LM ("KN" column) with the RFLM ("RF" column).

Table 3: *Main Perplexity Results*

Model	Method	KN	RF
$P(W, S)$	tuple 3gm	358	306
	decomp.	274	**251**
$P(W)$ $= \sum_S P(W, S)$	tuple 3gm	69.3	67.2
	decomp.	66.8	**64.2**
$P(W)$	word 3gm	66.1	62.3

As shown in Table 3, the best perplexity resulted from the decomposition method using the RFLM in both the model $P(W, S)$, where the prosodic breaks were given, and the model $P(W) = \sum_S P(W, S)$, where the prosodic breaks were hidden.

If we knew nothing about the prosodic breaks, we could still build a trigram LM with the modified Kneser-Ney smoothing or the RF. We called it the "word 3gm" method and put the perplexity results in the last row of Table 3. We observed that although our best number for the model $P(W) = \sum_S P(W, S)$ was better than a modified Kneser-Ney-smoothed trigram LM, it was outperformed by the basic RFLM, as shown in the bottom right corner of Table 3. The reason was that in the model $P(W) = \sum_S P(W, S)$, we were trying to predict a prosodic break from its proceeding words and breaks, which correlated poorly with it, instead of from its corresponding acoustic features.

Therefore we concluded that given prosodic breaks, we could successfully reduce the LM perplexity by a significant margin with the RFLM and the decomposition formula (5).

5. Discussion

Given that we could build a good LM when the prosodic breaks were provided (Table 2) but could not when they were not (Table 3 model $P(W) = \sum_S P(W, S)$), it is clear that we should get the prosodic breaks from the acoustics, instead of predicting them from words. In fact, the prediction of prosodic breaks from words was so bad that it killed the gain we had from using them to improve the word prediction. Therefore we propose the following procedure of using prosodic breaks in an ASR system:

- Generate an N-best list of hypotheses from a standard ASR system;
- For each hypothesis, align the words with the acoustics using the Viterbi algorithm; find out the regions between words and predict their prosodic breaks from the acoustic features using a prosody classifier;
- Rescore the N-best list with the model $\prod_i P(w_i|w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$.

Note that the model $\prod_i P(w_i|w_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$ is not a "pure" LM anymore since the s_i's come from the acoustics. However, because the acoustic features used to predict the breaks are different from those used to predict the words, we expect the new information would help choose a better hypothesis through the prosodically-informed LM.

6. Conclusions

We have presented our method that uses the RFLM to build LMs strengthened by prosodic break information. We showed that the ToBI break indices were not fine-grained enough for the task of language modeling. Using quantized posterior probabilities from a decision tree classifier as fine-grained prosodic breaks, we could reduce the perplexity by a significant margin. We also demonstrated that the RFLM was an ideal framework for incorporating various information like prosodic breaks into the existing LM in a principled way.

7. Acknowledgments

We would like to thank Zak Shafran, Markus Dreyer and the whole CLSP Workshop'05 PSSED team for preparing and sharing the data.

8. References

[1] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic

detection of sentence boundaries and disfluencies based on recognized words," in *Proceedings of ICSLP-1998*, vol. 5, 1998, pp. 2247–2250.

[2] J. Hirschberg and C. H. Nakatani, "Acoustic indicators of topic segmentation," in *Proceedings of ICSLP-1998*, 1998.

[3] J. Hale, I. Shafran, L. Yung, B. Dorr, M. Harper, A. Krasnyanskaya, M. Lease, Y. Liu, B. Roark, M. Snover, and R. Stewart, "PCFGs with syntactic and prosodic indicators of speech repairs with syntactic and prosodic indicators of speech repairsctic and prosodic indicators of speech repairs," in *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, 2006, pp. 161–168.

[4] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," *Computer Speech and Language*, vol. 17, no. 4, pp. 311–328, 2003.

[5] K. Chen, S. Borys, M. Hasegawa-Johnson, and J. Cole, "Prosody dependent speech recognition with explicit duration modelling at intonatinal phrase boundaries," in *Proceedings of INTERSPEECH-2003*, 2003, pp. 393–396.

[6] J. E. Fosler-Lussier, "Dynamic pronunciation models for automatic speech recognition," Ph.D. dissertation, University of California, Berkeley, CA, USA, 1999.

[7] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proceedings of Eurospeech-1999*, 1999.

[8] K. Hirose, N. Minematsu, and M. Terao, "Statistical langage modeling with prosodic boundaries and its use for continuous speech recognition," in *Proceedings of ICSLP-2002*, 2002.

[9] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 2003, pp. 147–154.

[10] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," in *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 119–121.

[11] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Proceedings of ICSLP-1992*, 1992, pp. 867–870.

[12] P. Xu and F. Jelinek, "Random forests in language modeling," in *Proceedings of EMNLP 2004*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 325–332.

[13] P. Xu and L. Mangu, "Using random forest language models in the IBM RT-04 CTS system," in *Proceedings of INTERSPEECH-2005*, 2005, pp. 741–744.

[14] Y. Su, F. Jelinek, and S. Khudanpur, "Large-scale random forest language models for speech recognition," in *Proceedings of INTERSPEECH-2007*, vol. 1, 2007, pp. 598–601.

[15] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modelling," *Computer Speech and Language*, vol. 10, pp. 187–228, 1996.

[16] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in lanugage modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 355–372, 2000.

[17] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech understanding: An overview of recent research at sri," in *Proceedings of ISCA Workshop on Speech Recognition and Understanding*, 2001, pp. 13–16.

[18] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proceedings of ICASSP-2003*, 2003.

[19] M. Dreyer and I. Shafran, "Exploiting prosody for PCFGs with latent annotations," in *Proceedings of INTERSPEECH-2007*, 2007.

[20] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue," *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[21] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[22] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 283–332, 2000.

[23] E. Charniak, "Immediate-head parsing for language models," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 124–131.

[24] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of HLT/NAACL 2003*, 2003, pp. 4–6.

[25] K. Duh and K. Kirchhoff, "Automatic learning of language model structure," in *Proceedings of COLING-2004*, 2004.

[26] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "A tree-based statistical language model for natural language speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 1001–1008, 1989.

[27] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, pp. 359–394, 1999.