

Using Prosodic Features in Language Models for Meetings

Songfang Huang and Steve Renals

The Centre for Speech Technology Research
University of Edinburgh
Edinburgh, EH8 9LW, UK
{s.f.huang, s.renals}@ed.ac.uk

Abstract. Prosody has been actively studied as an important knowledge source for speech recognition and understanding. In this paper, we are concerned with the question of exploiting prosody for language models to aid automatic speech recognition in the context of meetings. Using an automatic syllable detection algorithm, the syllable-based prosodic features are extracted to form the prosodic representation for each word. Two modeling approaches are then investigated. One is based on a factored language model, which directly uses the prosodic representation and treats it as a ‘word’. Instead of direct association, the second approach provides a richer probabilistic structure within a hierarchical Bayesian framework by introducing an intermediate latent variable to represent similar prosodic patterns shared by groups of words. Four-fold cross-validation experiments on the ICSI Meeting Corpus show that exploiting prosody for language modeling can significantly reduce the perplexity, and also have marginal reductions in word error rate.

1 Introduction

Prosody has long been studied as a knowledge source for speech understanding, and has been successfully used for a variety of tasks, including topic segmentation [1], disfluency detection [2], speaker verification [3], and speech recognition [4–6].

Recently there has been an increasing research interest in multiparty conversations, such as group meetings. Speech in meetings is more natural and spontaneous than read or acted speech. The prosodic behaviours for speech in meetings are therefore much less regular. Can prosody aid the automatic processing of multiparty meetings? Shriberg et al. [2] gave the answer ‘yes’ to this question, from the evidence of successfully exploiting prosodic features for predicting punctuation, disfluencies, and overlappings in meetings. It has also been noted that prosodic features can serve as an efficient non-lexical feature stream for tasks such as dialogue acts (DA) segmentation and classification, speech summarization, and topic segmentation and classification in the meetings domain.

This paper is concerned with the question of exploiting prosody to aid automatic speech recognition (ASR) in the context of meetings. Three essential components in a state-of-the-art ASR system, namely the acoustic model, language

model (LM), and lexicon, can all potentially serve to accommodate prosodic features. In this paper we are interested in exploiting prosodic features in language models for ASR in meetings.

The goal of a language model is to provide a predictive probability distribution for the next word conditioned on the strings seen so far, i.e., the immediately preceding $n - 1$ words in a conventional n -gram model. In addition to the previous words, prosodic information associated with the audio stream, which is parallel to the word stream, can act as a complementary knowledge source for predicting words in LMs. This understanding is the initial motivation for this work.

Due to the large vocabulary size in LMs (typically greater than 10,000 words), incorporating prosodic information in language models is more difficult than in other situations such as DA classification which has a much smaller number of target classes (typically several tens). To exploit prosody for LMs, a central question is how the relationship between prosodic features F and the word types W , $P(W|F)$, may be modeled. In this paper, two models will be investigated, namely the factored language model (FLM) [7] and the hierarchical Bayesian model (HBM) [8]. In the FLM-based approach, conditional probabilities $P(W|F)$ are directly estimated from the co-occurrences of words and prosody features via maximum likelihood estimation (MLE). The HBM-based approach provides a richer probabilistic structure by introducing an intermediate latent variable—in place of a direct association between words and prosodic features—to represent similar prosodic patterns shared by groups of words. This work is characterised by an automatic and unsupervised modeling of prosodic features for LMs in two senses. First, the prosodic features, which are syllable-based, are automatically extracted from audio. Second, the association of words and prosodic features is learned in an unsupervised way.

The rest of this paper is organized as follows. The next section reviews some related work on exploiting prosody for ASR. The ICSI Meeting Corpus, used throughout this paper, is described in Sect.3. The extraction of prosodic features is discussed in Sect.4. Section 5 focuses on the modeling approaches, including FLM-based and HBM-based methods. Experiments and results are reported in Sect.6, followed by a discussion in the final section.

2 Related Work

It is well accepted that humans are able to understand prosodic structure without lexical cues. Sub-lexical prosodic analysis [9] attempts to mimic this human ability using syllable finding algorithms based on band pass energy. Prosodic features are then extracted at the syllable level. The extraction of syllable-based prosodic features is attractive, because the syllable is accepted as a means of structuring prosodic information. This approach was verified on DA and hotspot categorization [9], which encourages us to utilize syllable-based prosodic features in LMs for ASR.

A basic approach to incorporate prosodic features in acoustic models for ASR uses “early integration”, in which the prosodic features are appended to the standard acoustic features [10]. Early work to utilize prosody in language models used prosodic features to evaluate possible parses for recognized words, which in turn would be the basis for reordering word hypotheses [11]. More recently, approaches that integrate prosodic features with LMs have emerged, in which LMs are conditioned on prosodic evidence by introducing intermediate categories. Taylor et al. [12] took the dialogue act types of utterances as the intermediate level, by first using prosodic cues to predict the DA type for an utterance and then using a DA-specific LM to constrain word recognition. Stolcke et al. [4] instead used prosodic cues to predict the hidden event types (filled pause, repetition, deletion, repair) at each word boundary with hidden event n -gram model, and then conditioned the word portion of the n -gram on those hidden events. Chan et al. [6] proposed to incorporate prosody into LMs using maximum entropy. However, the prosodic features they used were derived from manual ToBI transcriptions. An example of using prosody in the lexicon was provided by Chen et al. [5], where prosodic features, such as stress and phrase boundary, were included in the vocabulary. Each word had different variations corresponding to stress and whether or not it precedes a prosodic phrase boundary. This approach attempted to capture the effects of how prosodic features affect the spectral properties of the speech signal and the co-occurrence statistics of words.

Most research on using prosodic features for ASR has been applied to small and task-oriented databases. The goal of effectively using prosody for large-vocabulary speech recognition, such as recognition of meeting speech, still remains elusive. There has been little work in this direction in the meeting domain. One reason for this is due to the difficulty of modeling the relationship between symbolic words and normally non-symbolic prosodic features. Therefore, to find an approximate prosodic representation for each word in the vocabulary is one way to use prosodic features for ASR.

Realizing the difficulty of modeling prosody via intermediate representations, Shriberg et al. proposed direct modeling of prosodic features [13]. In this approach, prosodic features are extracted directly from the speech signal. Machine learning techniques (such as Gaussian Mixture Models, and decision trees) then determine a statistical model to use prosodic features in predicting the target classes of interest. No human annotation of prosodic events is required in this case. However, using prosodic features to predict very large number of target categories like words will again fail in capturing the prosodic discriminabilities.

3 Meeting Corpus

The experiments reported here were performed using the ICSI Meeting Corpus [14], which is a corpus of 75 naturally-occurring, unrestricted, and fairly unstructured research group meetings, each averaging about an hour in length. We performed our experiments using a four-fold cross-validation procedure in which

we trained on 75% of the data and tested on the remaining 25%, rotating until all the data was tested on. The corpus was divided into four folds, first by ordering all the sentences in sequence, and then for each fold sequentially selecting every fourth sentence. After further removing the sentences that are too short in length to extract prosodic features, this procedure resulted in the data set summarised in Table 1.

Table 1. The summary of the four-fold cross-validation setup on the ICSI Meeting Corpus used in this paper.

Fold	Number of Sentences	Number of Tokens
0-fold	27,985	209,766
1-fold	27,981	208,554
2-fold	27,968	208,294
3-fold	27,975	205,944

4 Prosodic Feature

A notable aspect of the prosodic features used here is that they are syllable-based. It is reasonable to address prosodic structures at the syllable level, because prosodic features relating to the syllable reflect more clearly perceptions of accent, stress and prominence. The syllable segments were automatically detected based solely on the parallel acoustic signals using an automatic syllable detection algorithm. The framework for the extraction of syllable-based prosodic features is shown in Fig.1, which follows an approach to automatic syllable detection suggested by Howitt [15], which in turn was originated in work by Mermelstein [16].

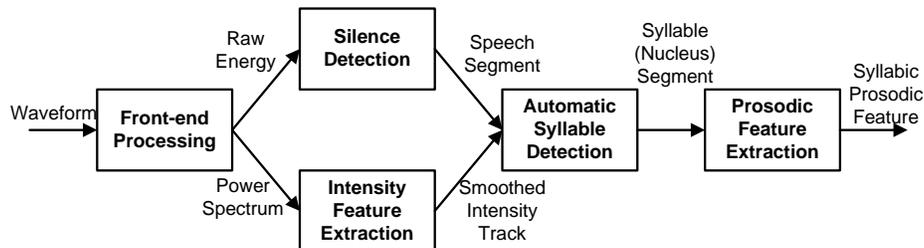


Fig. 1. The framework of the extraction of syllable-based prosodic features.

1. **Front-end Processing** The speech signal was first framed using a 16 ms Hamming window with a shift period of 10 ms. The raw energy before windowing and pre-emphasis was computed for each frame and saved in log magnitude representation for subsequent silence detection. A 256-point FFT was used to compute the power spectrum.
2. **Silence Detection** The raw energy data was smoothed using a 6th-order low-pass 50 Hz filter. Each frame was classified into either speech or silence based solely on whether or not the log frame energy was above a threshold. A running window consisting of 10 consecutive frames was used to detect the onsets of speech and silence. The detected speech segments, which were further extended by 5 frames at both sides, were fed into the following syllable detection.
3. **Intensity Feature Extraction** A single measure of intensity was computed, following Howitt’s adjusted features [15]. A 300–900 Hz band-pass filter was used to filter out energy not belonging to vowels. By a weighted summation (converted to magnitude squared forms) of the spectral bins within 300–900 Hz frequencies from the spectrogram, an intensity track (converted back to decibels) was computed for syllable detection, which again was smoothed by a low-pass 50 Hz filter to help reduce small peaks and noise.
4. **Automatic Syllable Detection** The recursive convex hull algorithm [16], which is a straightforward and reliable syllable detection algorithm, was used to find the nuclei by detecting peaks and dips in the intensity track computed in the above step. The syllables were then obtained by extending the nuclei on both sides, until a silence or a boundary of adjacent nuclei is detected.
5. **Prosodic Feature Extraction** Four prosodic features were extracted for each syllable consisting of the duration of syllable, the average energy, the average F0, and the slope of F0 contour. F0 information was obtained using the ESPS `get_f0` program.

We ran vector quantization (VQ), with 16 codewords (labeled ‘s0’ to ‘s15’) over all the 892,911 observations of syllable-based prosodic features in the ICSI Meeting Corpus. Before running VQ, each feature was normalized to unit variance.

The syllables belonging to an individual word were obtained by aligning the word with the syllable stream according to a forced time alignment at the word level, and selecting those syllables whose centres were within the begin and end times of words. By concatenating relevant VQ indices for syllables, we obtained the symbolic representations of prosodic features at the word level, which can then serve as potential cues for language modeling. For example, the prosodic representation for word ‘ACTUALLY’ might be the symbol ‘s10s12s6’, or ‘s10s15s6’ in other contexts.

5 Modeling Approach

5.1 Factored Language Model

One straightforward method for modeling words and prosodic features is to use MLE based on the co-occurrences of words W and the prosodic representations

F , i.e., training a unigram model $P(W|F) = \frac{\text{Count}(F,W)}{\text{Count}(F)}$. This unigram model can then be interpolated with conventional n -gram models. More generally, we can use the FLM [7] to model words and prosody deterministically. The FLM, initially developed to address the language modeling problems faced by morphologically rich or inflected languages, is a generalization of standard n -gram language models, in which each word w_t is decomposed into a bundle of K word-related features (called *factors*), $w_t \equiv f_t^{1:K} = \{f_t^1, f_t^1, \dots, f_t^K\}$. Factors may include the word itself. Each word in an FLM is dependent not only on a single stream of its preceding words, but also on additional parallel streams of factors. Combining with interpolation or generalized parallel backoff (GPB) [7] strategies, multiple backoff paths may be used simultaneously. The FLM’s factored representation can potentially accommodate the multimodal cues, in addition to words, for language modeling—in this case the prosodic representations. This configuration allows more efficient and robust probability estimation for those rarely observed word n -grams.

Supposing the word w_t itself is one of the factors $\{f_t^1, f_t^1, \dots, f_t^K\}$, the joint probability distribution of a sequence of words (w_1, w_2, \dots, w_T) in FLMs can be represented as the formalism shown in (1), according to the chain rule of probability and the n -gram-like approximation.

$$\begin{aligned}
 P(w_1, w_2, \dots, w_T) &= P(f_1^{1:K}, f_2^{1:K}, \dots, f_T^{1:K}) \\
 &= \prod_{t=1}^T P(f_t^{1:K} | f_{t-1}^{1:K}, f_{t-2}^{1:K}, \dots, f_1^{1:K}) \\
 &\approx \prod_{t=1}^T P(w_t | f_{t-n+1:t-1}^{1:K}) \tag{1}
 \end{aligned}$$

There are two key steps to use FLMs. First an appropriate set of factor definitions must be chosen. We employed two factors: the word w_t itself and the syllable-based prosodic representation f_t , as shown in Fig.2(A). Second it is necessary to find the suitable FLM models (with appropriate model parameters and interpolation/GPB strategy) over those factors. Although this task can be described as an instance of the structure learning problem in graphical models, we heuristically designed the model structure for FLMs. It is convenient to regard this FLM-based model as an interpolation of two conventional n -gram models $P(w_t|w_{t-1}, w_{t-2})$ and $P(w_t|w_{t-1}, f_t)$:

$$P_{\text{FLM}}(w_t|w_{t-1}, w_{t-2}, f_t) = \lambda_{\text{FLM}}P(w_t|w_{t-1}, w_{t-2}) + (1 - \lambda_{\text{FLM}})P(w_t|w_{t-1}, f_t) \tag{2}$$

Figure 2(B) shows the parallel backoff graph used in the experiments for factors w_t and f_t . We perform the interpolation in a GPB framework, as depicted in Fig.2, manually forcing the backoff from $P(w_t|w_{t-1}, w_{t-2}, f_t)$ to two parallel paths by setting a very large value of *gmin* for $P(w_t|w_{t-1}, w_{t-2}, f_t)$.

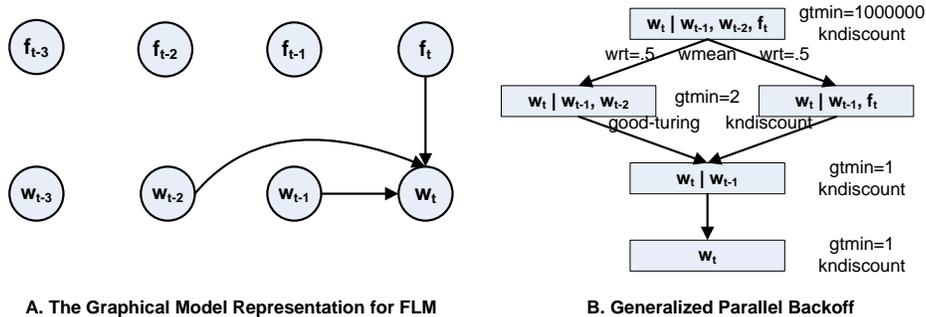


Fig. 2. (A) A directed graphical model representation for the factor configuration in a FLM over factors including words w_t , the prosodic representations f_t . (B) The generalized parallel backoff graph for w_t and f_t used in the experiments.

5.2 Hierarchical Bayesian Model

We argue that it is essential but difficult to find intermediate symbolic representations to associate words and low-level prosodic features for language modeling. In this paper, we have categorized syllable-based prosodic features into 16 classes, and represented the prosodic features for each word as a concatenation of indices for syllables belonging to that word. The FLM-based approach uses this prosodic information by directly associating word and prosodic representations. One limitation of this FLM-based approach is that there may be too many varieties of prosodic representations for individual words, due to the errors introduced by the automatic syllable detection and forced alignment. For example, the word ‘ABSOLUTELY’ in the ICSI Meeting Corpus has more than 100 different prosodic representations. Language models trained via MLE using such prosodic representations will be more likely to overfit to the training data. Rather than the direct association of words and prosodic representations, we introduce a latent variable between word and prosody and assume a generative model that generates words from prosodic representations through the latent variable. This probabilistic generative models is investigated within the framework of hierarchical Bayesian models [8].

Topic models have recently been proposed for document modeling to find the latent representation (*topic*) connecting documents and words. Latent Dirichlet allocation (LDA) [17] is one such topic model. LDA is a three-level hierarchical Bayesian model, in which each document is represented as a random mixture over latent topics, and each topic in turn is represented as a mixture over words. The topic mixture weights θ are drawn from a prior Dirichlet distribution:

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_K^{\alpha_K-1} \quad (3)$$

where $\alpha = \{\alpha_1, \dots, \alpha_K\}$ represents the prior observation count of the K latent topics with $\alpha_i > 0$. The LDA model is based on the “bag-of-words” assumption,

that is, words in a document exchangeably co-occur with each other according to their coherent semantic meanings. In this sense, LDA can be considered as a probabilistic latent semantic analysis model. However what if we assume that words in a document exchangeably co-occur with each other according to their coherent prosodic patterns? This is the intuition of our use of LDA for the probabilistic association of words and prosody, which we call the prosody-topic model.

In a prosody-topic model, a document in the corpus is composed by including all those words that have the same prosodic representation (i.e., ‘s10s12s6’). The prosodic representation is then served as the author of that document. If we apply LDA over this corpus, we can extract the latent ‘topics’ connecting words and prosodic representations. Each topic is expected to have coherent prosodic patterns. Considering our prosodic representations in this paper, for example, words in one individual topic are expected to have the same number of syllables whose pronunciations are similar. Unlike LDA, we need to explicitly retain the prosodic representations in the prosody-topic model. On the other hand, if we regard the prosodic representations as the ‘authors’ for corresponding documents, the prosody-topic model leads to the author-topic model [18], in which each document has only one unique author.

In short, the general idea of the prosody-topic model is that each prosodic representation is represented by a multinomial distribution over latent topics, and each topic is represented as a multinomial distribution over words. Prosody thus serves the same role as semantics, being the guideline to cluster co-occurring words in a document. The goal of a prosody-topic model is to learn the distribution of words for each topic, which therefore finds the latent representations association the word and prosodic representations. The graphical model for prosody-topic model is shown in Fig.3, and the generative process for each document d can be described as follows.

1. Select the unique prosodic representation (author) label f for document d .
2. Choose topic proportions $\theta|\{f, \theta_{1:F}\}$ for document d according to f , each $\theta_f \sim \text{Dirichlet}(\alpha)$
3. For each of the N_d words w_n in document d :
 - (a) Choose a topic $z_n|\theta \sim \text{Mult}(\theta)$.
 - (b) Choose a word $w_n|\{z_n, \phi_{1:K}\} \sim \text{Mult}(\phi_{z_n}), \phi_{z_n} \sim \text{Dirichlet}(\beta)$.

Since each document only has a single author, the probability of words w_t given prosodic representations f_t in a prosody-topic model can be easily obtained by integrating out latent topics, as shown in (4):

$$P_{\text{HBM}}(w_t|f_t) = \sum_{k=1}^K P(t_k|f_t)P(w_t|t_k) = \sum_{k=1}^K \theta_{f_t t_k} \phi_{t_k f_t} \quad (4)$$

where t_k is one of the K topics, $\theta_{f_t t_k}$ and $\phi_{t_k f_t}$ can be learned by approximate inference methods, such as variational EM or Gibbs sampling. This unigram-like probability can be interpolated with conventional n -gram models:

$$P_{\text{HBM}}(w_t|w_{t-1}, w_{t-2}, f_t) = \lambda_{\text{HBM}}P(w_t|w_{t-1}, w_{t-2}) + (1 - \lambda_{\text{HBM}})P_{\text{HBM}}(w_t|f_t) \quad (5)$$

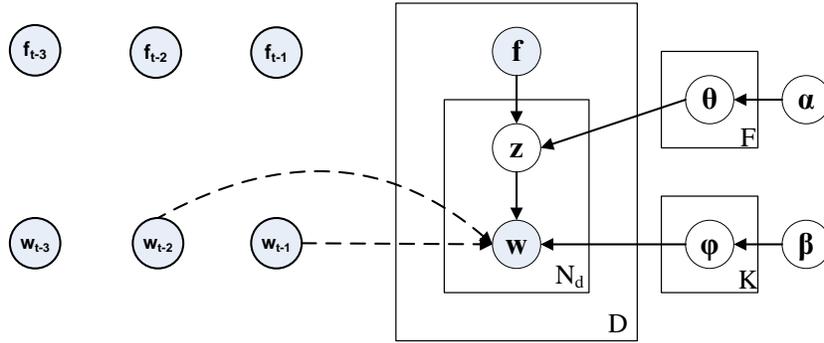


Fig. 3. The graphical model representation for the prosody-topic model (right) and its interaction with n -gram model (left), where shaded nodes denote observed random variables, while unshaded ones denote latent variables or parameters. The boxes are ‘plates’ representing the replications of a corresponding substructure.

6 Experiment and Result

We evaluated the FLM- and HBM-based approaches on the 4-fold cross-validation ICSI Meeting Corpus as described in Sect.3, in terms of perplexity (PPL) and word error rate (WER) respectively.

The FLM models were trained using the SRILM [19] toolkit¹, which has an extension for FLMs. Some modifications were made to the FLM toolkit regarding the manner of dealing with some special symbols such as ‘<s>’, ‘</s>’, and ‘NULL’, e.g., we manually set $P(w_t|w_{t-1}, w_{t-2}, \text{NULL}) = P(w_t|w_{t-1}, w_{t-2})$, and scored the end-of-sentence ‘</s>’ in perplexity calculations to account for the large number of short sentences in the meeting corpus. The FLM models share a common closed vocabulary of 50,000 word types with the AMI-ASR system [20]. The smoothing methods and parameters for FLM models are shown in Fig.2.

The prosody-topic models were trained using a publicly available Matlab topic modeling toolbox². The algorithm for inference is Gibbs sampling [21], a Markov chain Monte Carlo algorithm to sample from the posterior distribution. We chose the number of topics $K = 100$, and ran the Gibbs sampling algorithm for 2500 iterations, which took around one hour to finish the inference on a 3-fold ICSI data. Instead of automatically estimating the hyperparameters α and β , we fixed these two parameters to be $50/K$ and 0.01 respectively, as in [18].

The PPL results were obtained by successively testing on the specific fold with the language model trained on the other three folds. The interpolation weights λ_{FLM} and λ_{HBM} were both set to 0.5. Table 2 shows the PPL results on the 4-fold cross-validation ICSI Meeting Corpus. Both FLM-based and HBM-based approaches produce some reduction in PPL, especially the HBM-based approach

¹ <http://www.speech.sri.com/projects/srilm/>

² http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

has over 10% relative reduction in PPL than the baseline trigram model. One interesting thing we found during analysing the PPL results sentence-by-sentence is that those having higher probabilities than baseline trigrams normally have reasonable prosodic representations for words, i.e., representing the right number of syllables in a word.

Table 2. PPL results for 4-fold cross-validation experiments. **BASELINE-3G** denotes the baseline trigram results using the FLM toolkit. **FLM-3G-F** denotes the results for the FLM-based model, while **HBM-3G-F** for the HBM-based prosody-topic model.

TRAIN-TEST	BASELINE-3G	FLM-3G-F	HBM-3G-F
123 – 0	78.4	73.6	70.5
023 – 1	78.9	73.9	70.7
013 – 2	78.3	73.4	70.1
012 – 3	78.3	73.3	70.8
AVERAGE	78.5	73.5	70.5

Table 3 shows the WER results of n -best rescoring on the ICSI Meeting Corpus. It should be noted that the BASELINE-2G WER results were obtained during the first-pass decoding of the AMI-ASR system using an interpolated bigram LM trained on seven text corpora including Hub4, Switchboard, ICSI Meeting, and a large volume (around 1GB in size) of web data. The lattices were generated using this interpolated bigram LM. By retaining the time information for candidate words, the lattices were then used to produce n -best lists with time stamps for subsequent rescoring experiments via the *lattice-tool* program in the SRILM toolkit. In our experiments, the 500-best lists were produced from the lattices, which were then aligned with the syllable streams to get prosodic representation for each word, and finally reordered according to scores of different interpolated LMs to search for the best hypothesis. Marginal reductions in WER were observed in our experiments.

Table 3. Word error rate results, which share the same notations as in Table 2, except that the **BASELINE-2G** column represents the baseline results from the first-pass AMI-ASR decoding using an interpolated bigram model.

TRAIN-TEST	BASELINE-2G	BASELINE-3G	FLM-3G-F	HBM-3G-F
123–0	29.8	29.5	29.2	29.1
023–1	29.6	29.3	29.1	29.0
013–2	29.5	29.2	29.0	28.9
012–3	29.4	29.2	29.1	29.0
AVERAGE	29.6	29.3	29.1	29.0

7 Discussion and Future Work

In this paper we have investigated two unsupervised methods to exploit syllable-based prosodic features in language models for meetings. Experimental results on the ICSI Meeting Corpus showed our modeling approaches, both FLM-based and HBM-based, have significant reductions in PPL and marginal reductions in WER. The limited gains in WER may be partly caused by the following reasons. First, there are inevitably some errors in automatic syllable detection. It is hard for us to carry out evaluations on our syllable detection algorithm because of the lack of annotated data with syllable information. Second, additional errors are introduced by the forced alignment due to the overlapping cross-talk in meetings, which occasionally assigned an unreasonable number (i.e., more than 10) of syllables to a simple word. Third, the lattices were generated by an interpolated bigram model trained on a large corpus. This might prevent the recovery of more probable hypotheses from those n -best lists produced by a generalized LM, using specific LMs only trained on ICSI meeting data for rescoreing.

Considering the two modeling approaches, we are more interested in the HBM-based method. Bayesian language models [22], which provide an internally coherent probabilistic models and fit well in the hierarchical Bayesian model framework, have been proved to have comparable performance to the conventional n -gram models. In future, we will consider more tighter incorporation rather than simple interpolation, i.e., investigating the prosody-topic model and (Bayesian) language models in one united generative model within the hierarchical Bayesian framework. Moreover, meeting-specific cues will be taken into consideration for the prosody-topic model. For example, prosody encodes some information for DAs. DA in meetings normally has well-defined types. It is interesting to extend the prosody-topic model by investigating the relationship between word, prosody, and DA in one generative model.

Acknowledgement

We thank the AMI-ASR team for providing the lattices for rescoreing. This work is jointly supported by the Wolfson Microelectronics Scholarship and the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-243).

References

1. Shriberg, E., Stolcke, A., Hakkani-Tür, D.Z., Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* **32**(1-2) (2000) 127–154 Special Issue on Accessing Information in Spoken Audio.
2. Shriberg, E., Stolcke, A., Baron, D.: Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In: *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ (2001)

3. Sönmez, K., Shriberg, E., Heck, L., Weintraub, M.: Modeling dynamic prosodic variation for speaker verification. In: Proceedings of 5th International Conference on Spoken Language Processing, Sydney, Australia (1998) 3189–3192
4. Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G.: Modeling the prosody of hidden events for improved word recognition. In: Proceedings of 6th European Conference on Speech Communication and Technology, Budapest, Hungary (1999) 307–310
5. Chen, K., Hasegawa-Johnson, M.: Improving the robustness of prosody dependent language modelling based on prosody syntax dependence. In: IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands (2003) 435–440
6. Chan, O., Togneri, R.: Prosodic features for a maximum entropy language model. In: Proceedings of Interspeech (ICSLP'2006), Pittsburgh, US (2006)
7. Bilmes, J.A., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Proceedings of HLT/NACCL. (2003) 4–6
8. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. First edn. Chapman & Hall/CRC, London (1995)
9. Aylett, M.P.: Detecting high level dialog structure without lexical information. In: Proceedings of ICASSP'06, Toulouse, France (2006)
10. Lei, X., Siu, M., Hwang, M.Y., Ostendorf, M., Lee, T.: Improved tone modeling for mandarin broadcast news speech recognition. In: Proceedings of Interspeech (ICSLP'2006), Pittsburgh, US (2006)
11. Veilleux, N.M., Ostendorf, M.: Prosody/parse scoring and its applications in ATIS. In: Proceedings of ARPA HLT Workshop, Plainsboro, NJ (1993) 335–340
12. Taylor, P., King, S., Isard, S., Wright, H.: Intonation and dialog context as constraints for speech recognition. *Language and Speech* **41**(3-4) (1998) 489–508
13. Shriberg, E., Stolcke, A.: Direct modeling of prosody: An overview of application in automatic speech processing. In: Proceedings of International Conference on Speech Prosody, Nara, Japan (2004)
14. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proceedings of IEEE ICASSP, Hong Kong, China (2003)
15. Howitt, A.W.: Automatic Syllable Detection of Vowel Landmarks. PhD thesis, Massachusetts Institute of Technology (2000)
16. Mermelstein, P.: Automatic segmentation of speech into syllabic units. *Journal Acoustical Society of America* **58**(4) (1975) 880–883
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** (2003)
18. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada (2004)
19. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing, Denver, Colorado (2002)
20. Hain, T., Dines, J., Gaurau, G., Karafiat, M., Moore, D., Wan, V., Ordelman, R., Renals, S.: The development of the AMI system for the transcription of speech in meetings. In: Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Edinburgh, UK (2005)
21. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences, 101 (suppl. 1). (2004) 5228–5235
22. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the Annual Meeting of the ACL. Volume 44. (2006)