Latent Prosodic Modeling (LPM) for Speech with Applications in Recognizing Spontaneous Mandarin Speech with Disfluencies

Che-Kuang Lin and Lin-Shan Lee

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan
kimchy@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

In this paper, a new approach of Latent Prosodic Modeling (LPM) for analyzing the prosody of speech is presented. Based on a set of newly defined prosodic characters, prosodic terms, documents, and the Probabilistic Latent Semantic Analysis (PLSA) framework, prosody can be modeled using a set of prosodic states representing various latent factors such as speakers, speaking rate, utterance modality, intonation behavior, etc. in terms of some probabilistic relationships with the observed prosodic features. Organizing the training data based on this new model may also produce more delicate classification models for various speech processing applications considering the prosody. In the initial application example, we showed the use of this model on the task of disfluency IP detection for spontaneous Mandarin speech recognition, and improved IP detection accuracy and speech recognition performance were obtained in the experiments.

Index Terms: speech recognition, prosody, latent modeling, disfluency, Mandarin Chinese, spontaneous

1. Introduction

Prosodic information in speech signals is basically orthogonal to MFCC features and therefore should be useful for many spoken language processing applications [1,2,3,4]. However, very often such information was found useful in speech synthesis, but relatively difficult to use in speech recognition. The difficulties include, among many others, the fact that the prosody is usually speaker dependent [5], and that training corpora labeled with prosodic events usually require human efforts and are less available. In this paper, we try to develop a new framework of Latent Prosodic Model (LPM) for speech signals with a goal to at least handle parts of the above difficulties to a certain degree.

The concept of Latent Prosodic Modeling (LPM) is actually borrowed from the Probabilistic Latent Semantic Analysis (PLSA) very useful in the area of information retrieval [6]. In this approach, instead of directly counting the co-occurrence statistics between the document set $\{d_i\}$ and the term set $\{t_k\}$, a set of latent topics $\{z_l\}$ is created and the relationships between each document d_i and each term t_k are modeled by a probabilistic framework via these latent topics:

$$P(t_k \mid d_i) = \sum_{l=1}^{L} P(t_k \mid z_l) P(z_l \mid d_i) \ , \forall i,k \qquad (1)$$

where the probabilities were trained with EM algorithms by maximizing the total likelihood function:

$$L_T = \sum_{i=1}^{N} \sum_{k=1}^{N'} n(t_k, d_i) \log P(t_k \mid d_i) \ , \qquad (2)$$

and $n(t_k, d_i)$ denotes the frequency count of t_k in d_i, and N and N' are the total number of documents and terms respectively. In the Latent Prosodic Modeling (LPM) developed here, t_k, d_i,and z_l are to represent prosodic terms, prosodic documents, and the latent prosodic states respectively, as will be clear below.

Disfluencies, as one of the primary sources of ill-formness in spontaneous speech, pose difficult but important problems for spontaneous speech processing. Substantial work has been reported in this area [7,8,9,10]. The structure of disfluencies is usually considered to be decomposed into three regions: the reparandum, an optional editing term, and the resumption. The disfluency interruption point (IP) is the right edge of the reparandum and can be identified using prosodic information. The performance of spontaneous speech recognition can be improved by correct detection of the disfluency interruption point (IP). In this paper, we use the detection of disfluency IP as an initial example to show the possible applications of the proposed LPM.

Below in Section 2, we present the basic framework for LPM, while in Section 3, we describe the improved models for IP detection using LPM. Section 4 then gives the recognition approach incorporating the IP information. The experimental results are discussed in Section 5, and the concluding remarks finally made in Section 6.

2. Latent prosodic modeling (LPM) for speech

The dynamic behavior of speech prosody is affected by various latent factors, such as speakers, speaking rate, utterance modality, intonation behavior, etc., which leads to the significant variations in the observed prosodic features. The goal of LPM is to perform delicate analysis of the prosody by properly modeling the wide variety of prosodic features in terms of such latent factors, referred to as prosodic states here.

The prosodic feature vectors can first be extracted for phones, syllables, words, phrases, etc. Vector quantization (VQ) can then be used to label the feature vectors into discrete codewords, referred to as prosodic characters. The n-grams of these prosodic characters are then referred to as prosodic terms. The prosodic behavior of a certain part of the speech signal is then referred to as a prosodic document, composed of and characterized by the various prosodic terms included. The use of the term "document" is borrowed from PLSA and is used metaphorically here. All these are illustrated in Figure 1, in which three levels of prosodic documents were considered in this paper: segments, utterances, and speakers. The segments are parts of an utterance obtained from the best fitting piece-wise linear function for the pitch contour [3].

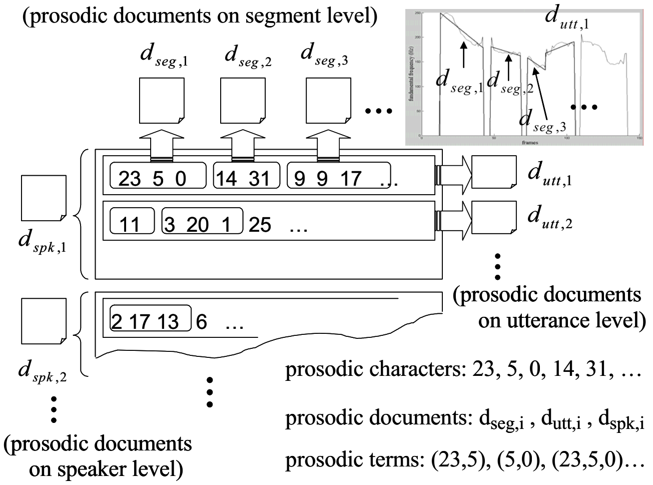
(prosodic documents on segment level)
$d_{seg,1}$ $d_{seg,2}$ $d_{seg,3}$ …
$d_{utt,1}$
$d_{seg,1}$ $d_{seg,2}$ $d_{seg,3}$ …
23 5 0 14 31 9 9 17 …
$d_{utt,1}$
11 3 20 1 25 …
$d_{utt,2}$
$d_{spk,1}$
2 17 13 6 …
(prosodic documents on utterance level)
$d_{spk,2}$
prosodic characters: 23, 5, 0, 14, 31, …
prosodic documents: $d_{seg,i}$, $d_{utt,i}$, $d_{spk,i}$
prosodic terms: (23,5), (5,0), (23,5,0)…
(prosodic documents on speaker level)

Figure 1 *Prosodic characters, terms and documents for latent prosodic modeling (LPM).*

For the set of each level of prosodic documents $\{d_i\}$ and the included prosodic terms $\{t_k\}$, we can then train a PLSA model as in equations (1) (2) by introducing a set of latent factors $\{z_l\}$, referred to as prosodic states here, and related all the prosodic terms t_k and prosodic documents d_i to the prosodic states z_l in terms of probabilistic distributions as shown in equation (1). This is the LPM proposed here in this paper. With such a model, the complicated behavior of the many prosodic features can be analyzed in terms of the latent prosodic states in some way. For instance, the similarity between any two prosodic documents d_i and d_j , $Sim_{LPM}(d_i,d_j)$, can be estimated by their probability distributions with respect to the various prosodic states, $P(z_l\,|\,d_i)$ and $P(z_l\,|\,d_j)$, with the expression below as one example:

$$Sim_{LPM}(d_i,d_j)=\frac{\sum_l P(z_l\,|\,d_i)P(z_l\,|\,d_j)}{\sqrt{\sum_l [P(z_l\,|\,d_i)]^2}\sqrt{\sum_l [P(z_l\,|\,d_j)]^2}}\;. \qquad (3)$$

Many other distance metrics can also be used, such as the Kullback-Leibler distance and Mahalanobis distance.

The above model can be useful in many applications such as prosodic behavior classification, for example, using the distance measure in equation (3). We may also realize delicate classification models that is adapted to, say, a specific speaker, a kind of utterance modality, or a particular intonation context, using other efficient classification algorithms (e.g. maximum entropy) but based on LPM in an unsupervised manner. As illustrated in Figure 2, we can actively select the desired training set for a specific testing condition by LPM at each level of prosodic documents, the segments, utterances or speakers. Taking the speaker level for example, the speaker-type model based on a subset of training data produced by the group of speakers with similar prosodic properties may be obtained in this way.

On the other hand, LPM can also be used in an alternative framework to learn the patterns for different classes of speech signals in a supervised manner, referred to as anchor modeling here. In this approach, the prosodic documents associated with each desired class were merged into a super-document representing the characteristics of this class, and LPM was then performed upon the set of super-documents. Thus the prosodic characteristics of each class anchor, in terms of the relationships with each prosodic state, can then be analyzed.

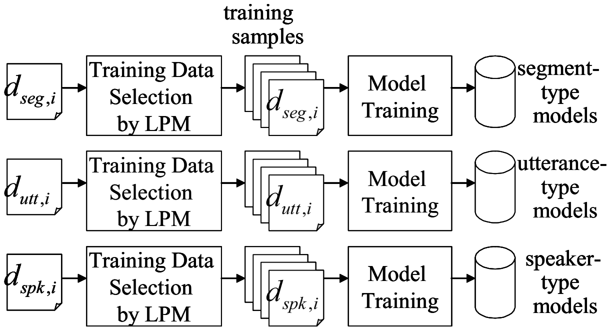

Figure 2 *Training of segment-, utterance-, speaker-type models based on Latent Prosodic Modeling (LPM).*

3. Interruption point (IP) detection in spontaneous Mandarin speech with LPM

We used IP detection for spontaneous Mandarin speech as an initial application example of the LPM proposed here. Due to the mono-syllabic structure of Chinese language, i.e., every character has its own meaning and is pronounced as a monosyllable, while a word is composed of one to several characters (or syllables), we extracted a whole set of prosodic features, pitch-related and duration-related, from every syllable boundary [11].

3.1. Maximum entropy (Maxent) modeling based on LPM

We successfully integrated the decision tree algorithm into the maximum entropy (maxent) model for IP detection based on the previous work [11], in which the feature functions for the maximum entropy model were derived using decision trees. This maxent model can be further improved by the LPM proposed here just as shown in Figure 2. The prosodic documents in the training corpus were first classified by LPM based on the latent prosodic states, and then more delicate maxent models based on the prosody of the segment types, utterance types and speaker types can be trained. As illustrated in Figure 3, the classification scores obtained by the three delicated maxent models based on segment types, utterance types and speaker types were then combined with the score by the maxent model without LPM via a support vector machine (SVM) with a radial basis kernel using the LIBSVM tool [12].

3.2. Anchor-based model with LPM

In this approach, we established with LPM a set of five anchors, each for one out of the four IP classes (overt repair, abandoned utterances, direct repetition, partial repetition) and the non-IP boundaries, to detect the disfluency IPs. As mentioned previously, prosodic documents in the training corpus associated with each of the above five classes were merged into five super-documents representing the prosodic characteristics of the four IP classes and non-IP, which produced a set of corresponding prosodic anchors after LPM. We similarly trained such anchor models on the three levels, i.e., for segment-, utterance- and speaker types, as in Figure 2, and combined the scores using SVM as in Figure 3.

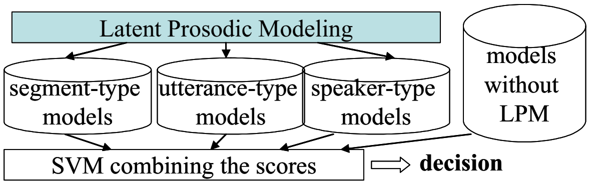

Figure 3 *Integration of LPM-based classification models with SVM.*

3.3. LPM-based feature expansion for Maxent

In addition, the probabilities that each prosodic state z_l is related to the prosodic document d_i, $\{P(z_l \mid d_i), \forall l\}$, and the likelihood of the prosodic terms given the prosodic document $\{\prod_{t_k \in d_i} P(t_k \mid d_i) = \prod_{t_k \in d_i} \sum_{l=1}^{L} P(t_k \mid z_l) P(z_l \mid d_i)\}$, obtained from LPM for prosodic documents at each level, can also be directly used as another set of features, together with other prosodic features for the maxent models.

4. Speech recognition with IP detection

Here we present the way to incorporate the IP detection results into the speech recognition processes. The IP detection gave the probability for each syllable boundary to be an IP (very often zero) along a sequence of word hypotheses. For each utterance, we combined such information for each path in an n-best list, where the probability for each syllable boundary to be an IP was the weighted sum over all paths in the n-best list, using the total likelihood scores for the paths as the weights. This gave each syllable boundary a final probability to be an IP, which was used in the following search process over the word graph.

We rescored the word graph based on the maximum a posterior (MAP) principle considering the prosodic information:

$$
\begin{aligned}
W^* &\equiv \arg\max_W P(W \mid X, F) \\
&= \arg\max_W P(W \mid F) P(X \mid W, F) \\
&\cong \arg\max_W P(W \mid F) P(X \mid W)
\end{aligned}
\tag{4}
$$

where X and F are the acoustic and prosodic feature sequences respectively, the recognized word sequence W^* is the one which maximizes the posterior probability P(W|X,F), and the last expression was based on the assumption that X and F can be approximated as independent given the word sequence W. P(W|F) is modeled considering the probabilities for the different disfluency IP classes as follows:

$$
\begin{aligned}
P(W \mid F) &= \prod_n P(w_n \mid w_{n-N+1}^{n-1}, F) \\
&= \prod_n \sum_c P(c \mid w_{n-N+1}^{n-1}, F)^{\lambda} P(w_n \mid w_{n-N+1}^{n-1}, c) \ ,
\end{aligned}
\tag{5}
$$

where w_{n-N+1}^{n-1} is the N-1 words before the n-th word w_n, c is one out of the five IP classes including non-IP, λ is a weight parameter, and $P(c \mid w_{n-N+1}^{n-1}, F)$ is approximated using the probability obtained from IP detection, or $P(c \mid w_{n-N+1}^{n-1}, F) \cong P(c \mid F)$. The word n-grams crossing different classes of IP boundaries (i.e. $P(w_n \mid w_{n-N+1}^{n-1}, c)$) were evaluated from disfluency corpus separately, and then interpolated with the baseline language model.

5. Experimental Results

5.1. Corpus

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) [13], collected from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. 8 dialogues out of the 30, with a total length of 8 hrs, produced by nine female and seven male speakers, annotated by adopting a taxonomy scheme of groups of spontaneous speech phenomena, were used in this research. The testing set was chosen to cover all the speakers. Table 1 summarizes the data used in the following experiments. Only 3.7% and 3.9% of the syllable boundaries are IPs.

Table 1. *The summary of experiment data.*

	train(6.9hr)	test(1.3hr)
Number of IPs / non-IPs	3432/89891	673/16529
Chance of non-IPs	96.3%	96.1%

5.2. IP detection with LPM-based models

Due to the limited quantity of the training data, we actually merged the four classes of IP into one and considered IP detection as a two-class classification problem in the experiments. For each syllable boundary, a decision between "non-IP" vs. "IP" was made with a probability. Figure 4(a) compares IP detection accuracies obtained using the delicate utterance-type LPM-based models as shown in Figure 2 to those using plain maxent and anchor models, with the training data for the delicate model selected using hierarchical agglomerative clustering (HAC) and k-nearest-neighbor (kNN) approaches. We see that kNN-based approach is better and the delicate utterance-type LPM-based approach apparently improved the performance.

Figure 4(b) then demonstrates the results when the delicate models of individual segment-, utterance- and speaker-types based on LPM were used, as well as all of them used together, all kNN-based. So the first and third bars in Figure 4(b) for each case are the same as those in Figure 4(a). The improvements obtainable from LPM are obvious at different levels especially for the anchor model, and the use of all the three levels is clearly better. So the prosodic information from different levels is complementary. The relatively lower performance for the segment-type model may be due to the relatively poor segmentation by the pitch contours.

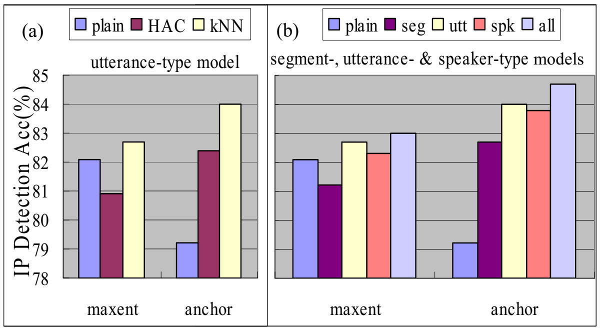

Figure 4. *IP detection accuracy using LPM-based maxent or anchor models, (a) with and without the utterance-type delicate models with training data selected by HAC- and kNN-based approaches. (b) with segment-, utterance-, and speaker-type delicate models with training data selected by kNN-based approach.*

5.3. LPM-based feature expansion for Maxent

As mentioned in section 3.3, LPM parameters $\{P(z_l \mid d_i)\}$ and $\{\prod P(t_k \mid d_i)\}$ can be used as extra features for the maxent model. The results for such case are shown in Figure 5, where the bar (a) is the same as the last bar for maxent in Figure 4(b), and the bars (b)(c)(d) are the results when these features were added individually and together. We can see that the expanded features are indeed useful. The last bar (e) is the result when the finally enhanced maxent model is combined with the anchor model by SVM, which eventually yielded the best result.

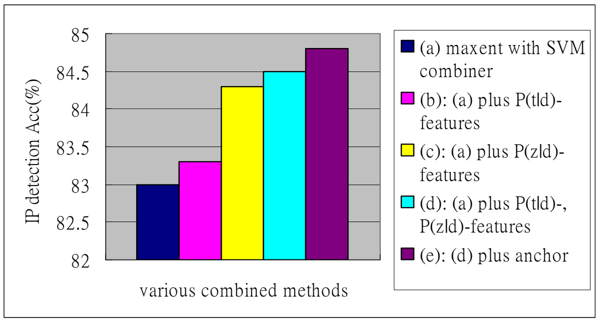

Figure 5 *The performance of the maxent model with expanded LPM-based features and finally integrated with the anchor model.*

5.4. Speech recognition results

The recognition experiments were performed with a lexicon of 50K entries, a trigram language model, and an intra-syllable right context dependent Initial/Final acoustic model set (a Mandarin syllable was decomposed into two parts: Initial and Final). Figure 6 shows the character accuracy with IP detection results considered as a function of the weight parameters λ in equation (5), using the formula described in Section 4 in the rescoring process, compared to the baseline without the disfluency information. We see that the highest improvement achievable with the LPM-based IP probability is about 2% of character accuracy when λ is about 0.9.

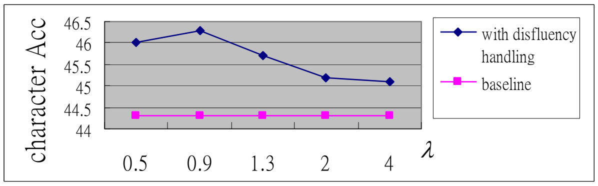

Figure 6 *Character accuracy with disfluency IP detection.*

6. Conclusions

We presented a new approach of modeling prosodic information in speech, LPM, and an initial application in spontaneous Mandarin speech recognition with disfluency IP detection. The LPM is a general approach for dealing with prosodic variation considering the latent factors not directly observable in speech signals. Experimental results showed improved performance when maxent and anchor models were enhanced by LPM. The results also verified the benefit of embedding disfluency information in the recognizer.

7. References

[1] Hirose, K., et al., “Use of prosodic features for speech recognition”, in Proc. of ICSLP, 2004.

[2] Vergyri, D., et al., “Prosodic knowledge source for automatic speech recognition”, in Proc. ICASSP, vol. 1, 2003, pp. 208-211.

[3] Shriberg, E., et al., “Prosody-based automatic segmentation of speech into sentences and topics”, Speech Communication, pp. 127-154, 2000.

[4] Chen, K., et al., “Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries”, in Proc. Eurospeech 2003.

[5] Chen, Z.-H., et al., “Probabilistic latent prosody analysis for robust speaker verification”, in Proc. of ICASSP, 2006.

[6] Hofmann, T., “Probabilistic latent semantic analysis”, Uncertainty in Artificial Intelligence, 1999.

[7] Lickley, R.J., “Juncture cues to disfluency”, in Proc. ICSLP, 1996

[8] Lendvai, P., et al., “Memory-based disfluency chunking”, in Proc. of DISS’03, pp. 63-66.

[9] Honal, M., et al., “Automatic disfluency removal on recognized spontaneous speech – rapid adaptation to speaker-dependent disfluencies”, in Proc. of ICASSP, 2005.

[10] Liu, Y., et al., “Comparing HMM, maximum entropy, and conditional random fields for disfluency detection”, in Proc. of Interspeech, 2005.

[11] Lin, C.-K. et al., “Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features”, in Proc. Interspeech 2005.

[12] Chang, C.-C. and Lin, C.-J., LIBSVM: A library for support vector machines, 2001. www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] Tseng, S.-C. 2004. Processing Spoken Mandarin Corpora. *Traitement automatique des langues*. Special Issue: Spoken Corpus Processing. 45(2): 89-108.