Linguistic Patterns in Spontaneous Speech, 193-211
2009-1-025-009-000271-1

Spontaneous Mandarin Speech Recognition with Disfluencies Detected by Latent Prosodic Modeling (LPM)

Che-Kuang Lin[1], Shu-Chuan Tseng[2], and Lin-Shan Lee[1]
National Taiwan University[1]
Academia Sinica[2]

In this paper, a new approach for improved spontaneous Mandarin speech recognition using Latent Prosodic Modeling (LPM) for disfluency interruption point (IP) detection is presented. The basic idea is to detect the disfluency interruption points (IPs) prior to the recognition, and then to incorporate these information into the recognition process via the second pass rescoring. For accurate detection of disfluency interruption points (IPs), prosodic information from local to global, from observable to latent, were integrated using the proposed Latent Prosodic Modeling (LPM). A whole set of new features were first defined for each syllable boundary obtained in the first pass recognition by carefully considering the special characteristics of Mandarin Chinese, and the importance of each feature with respect to each disfluency type was analyzed. Then, a set of prosodic characters, prosodic terms, and prosodic documents were defined to be used in the Probabilistic Latent Semantic Analysis (PLSA), based on which the 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. Using all these different levels of information, the approach of incorporating the decision tree into the maximum entropy model training was developed to enhance the IP detection accuracy. Experimental results indicated that the proposed set of features and the IP detection approach based on Latent Prosodic Modeling (LPM) were very useful, and the obtained information about disfluency actually benefited the speech recognition performance.

Key words: spontaneous speech recognition, disfluency, prosody, latent modeling, Mandarin Chinese

1. Introduction

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 (Lickley 1996, Lendvai et al. 2003, Honal & Schultz 2005, Liu et al. 2005). While analyses regarding different aspects of disfluency phenomena

have been conducted and much insight into the related acoustic and prosodic properties has been gained so far, we are trying to integrate all the valuable knowledge into a system for enhanced speech recognition. 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 is the right edge of the reparandum. The purpose of the research presented in this paper is to identify useful and important features in automatic detection of such disfluency interruption point (IP) in spontaneous Mandarin speech, and analyze how these features are helpful in speech recognition. The disfluencies considered here in this paper include the following four categories:

(i) Direct repetitions: the speaker repeats words in a way that can not be justified by grammatical rules. Many other cases of repetitions in Mandarin Chinese are perfectly legal syntactic constructions for emphasis purposes and so on, which should be excluded from this study.
(ii) Partial repetitions: only part of a word (including compound words) is repeated.
(iii) Overt repairs: the speaker modifies expressed words within an utterance.
(iv) Abandoned utterances: the speaker abandons an utterance and starts over.

Consider the following example in Mandarin Chinese:

是	進口	嗯	出口	嗎
shi4	**jin4kou3**	**EN**	**chu1kou3**	**ma1?**
is	import	[discourse particle]	export	[interrogative particle]

(*Do you import * uhn export products?*)

In this example, "uhn" is a filled pause and "export" is meant to correct "import", which is an overt repair. Here '*' denotes the right edge of the reparandum region, or the interruption point (IP) to be detected and utilized in recognition here. Consider another example:

因爲	因爲	他	有	健身	中心
yin1wei4	**yin1wei4**	**ta1**	**you3**	**jian4shen1**	**zhong1xin1**
because	because	it	has	fitness	center

(*Because * because it has a fitness center.*)

Here the speaker repeats the word "because" to restart the sentence.

Prosodic information in speech signals can be considered to be more or less orthogonal to MFCC features and therefore should be useful for many spoken language

processing applications (Hirose & Minematsu 2004, Vergyri et al. 2003, Shriberg et al. 2000, Chen et al. 2003). 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 (Chen et al. 2006), 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 (Hofmann 1999). 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:

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

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

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

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.

Below, we first describe the corpus used in this research in §2 and then introduce the set of proposed prosodic features as well as IP detection models in §§3 & 4. Then in §5, we present the basic framework for LPM, while in §6, we describe the improved models for IP detection using LPM. Section 7 then gives the recognition approach incorporating the IP information. Analysis regarding the contribution of different features for detection of different types of disfluency IPs is presented in §8. The experimental results are then discussed in §9, and the concluding remarks finally made in §10.

2. Corpus used in the research

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) (Tseng 2004, website http://mmc.sinica.edu.tw), collected from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. This corpus includes 30 digitized conversational dialogues with a total length of 27 hours. 8 dialogues out of the 30, with a total length of 8 hrs, produced by nine female and seven male speakers, were annotated by adopting a taxonomy scheme of four groups of spontaneous speech phenomena: disfluencies, sociolinguistic phenomena, particular vocalization, and unintelligible or non-speech sounds. Disfluencies here include breaks, word fragment, overt repairs, direct repetitions, abandoned utterances, discourse particles, and markers. In this paper, we only deal with direct repetitions, partial repetitions, overt repairs and abandoned utterances. The 8 hrs of annotated dialogues as mentioned above were used in this research. Due to the mono-syllabic structure of the Chinese language, i.e., in Mandarin Chinese every character has its own meaning and is pronounced as a monosyllable, while a word is composed of one to several characters (or syllables), every syllable boundary is considered as a possible interruption point (IP) candidate in this research. 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%

3. Prosodic features

As mentioned above, due to the mono-syllabic structure of Chinese language, every syllable boundary is considered as a possible interruption point (IP) candidate in this research. We therefore tried to define a whole set of prosodic features for each IP candidate, or each syllable boundary, and use them to detect the IPs. Many prosodic features have been proposed and proved useful for such purposes (Shriberg et al. 2000, Liu et al. 2003), and it has been found (Liu et al. 2005) that it is important to identify better features. Because this research is focused on IP detection, we tried to identify some IP specific features. Moreover, considering the special feature of Mandarin Chinese including the tonal language nature, some acoustic phenomena for Mandarin spontaneous speech may be quite different from those in English. Such consideration was reflected here by constructing a new set of features.

3.1 Pitch-related features

Pitch information is typically less robust and more difficult to use (Shriberg et al. 2000). Pitch contour stylization method has thus been used, and smoothing out the "micro-intonation" and tracking errors was found helpful for English (Shriberg et al. 2000, Liu et al. 2003). For a tonal language such as Mandarin Chinese, however, such "micro-intonation" apparently carries tone or lexical information, and thus should not be removed, although some approaches of pitch contour smoothing are certainly needed.

In this research, we used Principal Component Analysis (PCA) for syllable-wise pitch contour smoothing, instead of piece-wise linear stylization. For each syllable, the pitch contour was decimated or interpolated to become a vector with fixed dimension. PCA was then performed on such training vectors. By choosing the principal components with the largest eigenvalues, we projected the fixed dimension vectors onto the subspace spanned by the principal components to obtain the smoothed version of the pitch contours. Various pitch-related features were then extracted from these smoothed pitch contours, such as the pitch reset for boundaries being considered and so on. Quite several syllable-wise pitch-related features found useful in tone recognition were also used here, such as the average value of normalized pitch within the syllable, the average of absolute value of pitch variation within the syllable, the maximum difference of normalized pitch within the syllable and so on, all evaluated for the syllable before and after the boundary being considered. A total of 54 such pitch-related features were considered (Lin & Lee 2005, Lin et al. 2005).

3.2 Duration-related features

Duration features such as pause and phone duration features have been used to describe prosodic continuity and preboundary lengthening (Shriberg et al. 2000, Liu et al. 2003). By carefully examining the characteristics of IPs in our corpus, we hypothesized that deviation from the normal speaking rhythmic structure is an important cue to disfluency IP detection. For example, relatively sudden, sharp, discontinuous changes in speaking rate were consistently observed across IPs. We also hypothesized that certain ways of integration of pause and syllable duration fluctuations are important characteristics of the rhythmic structure of speech. Considering these observations, we derived the following set of duration-related features to try to detect IPs.

We first computed the average and standard deviation of syllable duration over several syllables before and after the boundary being considered. Then we calculated the ratio of the former to the latter. The possible ranges for evaluating the above statistics included one, two, three syllables as well as extending to the nearest pauses on both

sides. Another group of duration-related features were generated by jointly considering the pause duration and the duration parameters of the syllables before or after the pause. The product of these two different duration parameters represented some integration of the two types of information. Alternatively, normalizing the syllable duration parameters by the duration of a nearby pause being considered may emphasize the fluctuations of these syllable duration parameters. Finally, a total of 38 such duration-related features were considered (Lin & Lee 2005, Lin et al. 2005).

4. Interruption point (IP) detection

The approaches used for IP detection are discussed in this section. The IP detection task is considered as a classification problem here in this research. For each syllable boundary, a decision between "non-IP" vs. various types of "IPs" was made. Because IPs were relatively rare events, we used ensemble sampling (Liu et al. 2004) on training data to equate the prior probabilities for different classes. This made the model trained more sensitive to any features that distinguish the classes.

4.1 Decision tree (DT) and maximum entropy (Maxent) model

In the first approach, we used decision trees to learn from data, and to make prediction while testing (Liu et al. 2003). The decision was then made according to the posterior probability of the leaf node where the test sample of the syllable boundary went to. In the second approach, we applied the maximum entropy model to make the decision (Berger et al. 1996). In this model, a feature is expressed by a binary feature function $f_i(x, y)$, in which x denotes the feature sets and y denotes the outcome. The final expression for p(y|x) in this model takes the following form:

$$(3) \quad p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_i \lambda_i f_i(x, y)\right]$$

where $Z(x)$ is a normalization term.

The maximum entropy model is estimated by finding the parameters λ_i for each feature function $f_i(x, y)$ with the constraint that the expected values of the various feature functions match the empirical averages in the training data. In our experiments we used the L-BFGS parameter estimation method with Gaussian-prior smoothing to avoid overfitting.

4.2 Integration of DT and Maxent

Considering the decision trees and maximum entropy model mentioned above, we can find that each of them has some advantages and limitations in dealing with the problem here. Decision Trees can handle real-valued features directly while Maxent is working in a discrete style. On the other hand, by carefully designing the feature function, maximum entropy model may make finer decision on a certain feature parameter, while for the decision trees the training process only uses binary partitioning on feature parameters to split the data. When the problem is not linearly separable, it might not be possible for the decision trees to find a good partition without growing too deep. But too deep trees often lead to overfitting and thus degrade the performance on the testing set.

Based on all the above considerations, we developed a new approach to integrate the decision trees and maximum entropy model together (Lin & Lee 2005). In this approach, we used decision trees built with training data to derive the feature functions for the maximum entropy model, hoping to have the advantages of both. We first trained a set of decision trees by ensemble downsampling of the training data. Instead of growing the optimized tree by cross validation, we chose Bayesian criterion to grow the tree, resulting in a set of much deeper and bushy trees. Each leaf of all the trees was then used as a feature function in Maxent. In other words, a feature function was assigned "1" if and only if the sample being considered went to the corresponding tree leaf. Otherwise the feature function was 0-valued. By having each sample traversing down to the leaves, we had the feature function values all decided. The training procedure was then the same as the original maximum entropy model. While testing, the complete procedure was the same as that of training stage and the pre-trained trees were used again. This approach is referred to as the integrated maximum entropy (integrated maxent) model hereafter in this paper.

5. 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.

We defined prosodic terms and documents from speech signal as the unit of prosody analysis. Although these units are not defined based on any known theory of prosody, they served as units of analysis when not much the spoken content but the signal itself is available in a bottom-up recognition framework. The prosodic feature vectors can

first be extracted for phones, syllables, words, phrases, etc. (for the present research, for each syllable boundary as mentioned in §3). 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 Fig. 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 (Shriberg et al. 2000).

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, $P(z_l \mid d_i)$ and $P(z_l \mid d_j)$, with the expression below as one example:

$$Sim_{LPM}(d_i, d_j) = \frac{\sum_l P(z_l \mid d_i) P(z_l \mid d_j)}{\sqrt{\sum_l [P(z_l \mid d_i)]^2} \sqrt{\sum_l [P(z_l \mid d_j)]^2}} \quad . \tag{4}$$

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

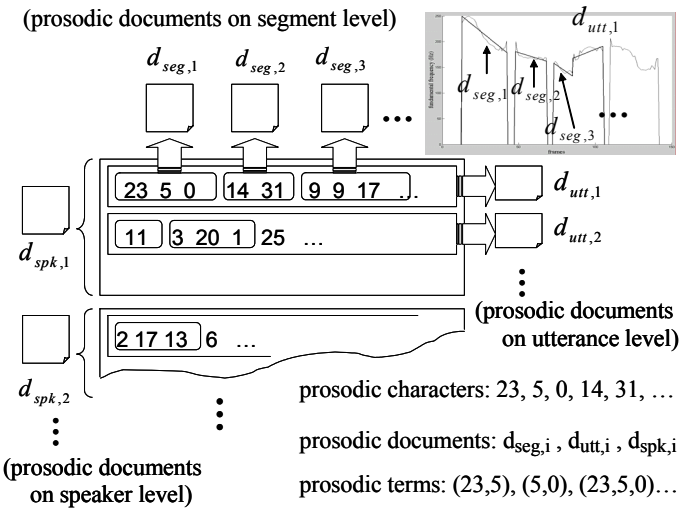

Figure 1: Prosodic characters, terms and documents for latent prosodic modeling (LPM)

Although below we use this model for IP detection (Lin & Lee 2006), the above model can also be useful in many applications such as prosodic behavior classification, for example using the distance measure in equation (4). 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. integrated maxent) but based on LPM in an unsupervised manner. As illustrated in Fig. 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.

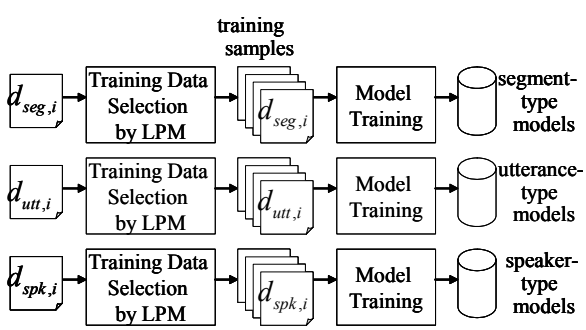

Figure 2: Training of segment-, utterance-, speaker-type models based on Latent Prosodic Modeling (LPM)

On the other hand, LPM can also be used in an alternative framework (Lin & Lee 2006) 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.

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

We used LPM proposed here for IP detection for spontaneous Mandarin speech.

6.1 Integrated maximum entropy (integrated maxent) modeling based on LPM

The integrated maxent model mentioned above can be further improved by the LPM proposed here just as shown in Fig. 2. The prosodic documents in the training corpus were first classified by LPM based on the latent prosodic states, and then more delicate integrated maxent models based on the prosody of the segment types, utterance types and speaker types can be trained. As illustrated in Fig. 3, the classification scores obtained by the three delicated integrated maxent models based on segment types, utterance types and speaker types were then combined with the score by the integrated maxent model without LPM via a support vector machine (SVM) with a radial basis kernel using the LIBSVM tool (Chang & Lin 2004).

6.2 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 Fig. 2, and combined the scores using SVM as in Fig. 3.

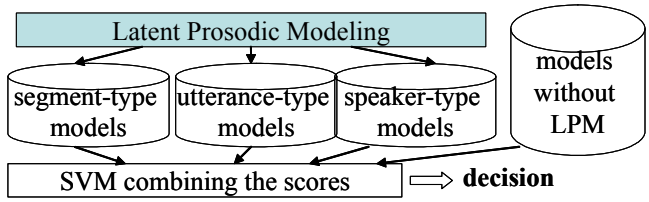

Figure 3: Integration of LPM-based classification models with SVM

6.3 LPM-based feature expansion for integrated 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 integrated maxent models.

7. 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 then 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:

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

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:

$$(6)\quad P(W \mid F) = \prod_{n} P(w_n \mid w_{n-N+1}^{n-1}, F)$$
$$= \prod_{n} \sum_{c} P(c \mid w_{n-N+1}^{n-1}, F)^{\lambda} P(w_n \mid w_{n-N+1}^{n-1}, c)$$

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.

8. Feature analysis for disfluency detection

8.1 Comparison between duration- and pitch- related features for different disfluency types

To get a further insight into the characteristics of various disfluency categories and the IP detection process, we tried to find the relation between the features used and the IP detection performance. A partial feature selection analysis was performed upon the full feature set mentioned earlier. In this approach, we excluded each single feature from the full set and then perform the complete IP detection process in each small experiment, to find out how much the IP detection performance was degraded due to the missing of this single feature. Here the performance is in terms of recall rate only. Because we grouped all the four types of disfluencies together into a single class due to the small size of the corpus, precision for each disfluency type was not obtainable, while recall was. The results discussed here were obtained from integrated maximum entropy model.

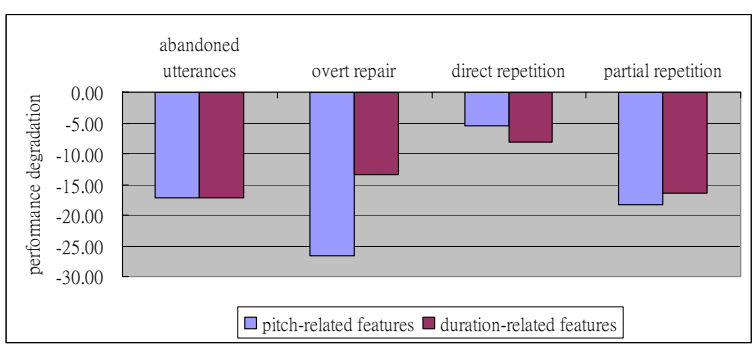

Figure 4: Performance degradation for the four disfluency types with respect to the two feature categories

First, to see how pitch-related and duration-related features contribute to the IP detection of different types of disfluencies, we compared the performance degradation for the four disfluency types being considered with respect to the two feature categories. In Fig. 4, we show the most serious performance degradation caused by removing one single feature from the two categories of either pitch-related or duration-related features. We can find that for overt repair and partial repetition, pitch-related features play relatively more important role for IP detection, and this is specially apparent for overt repair. This is in good consistency with the earlier findings (Tseng 2006a) that overt repairs are produced partly because the correction of the delivered information is required, and partly because the speaker changes his/her language planning. It is often true that when overt repairs are produced within utterances, the F0 level of the onset of the resumption part is approximately reset to that of the onset of the reparandum. In other words, the resumption part should fit seamlessly into the original utterance after removing the problematic items. Then the cleaned utterance should look like a natural utterance that obeys the normal F0 declination. In addition, intonation units have been defined and analyzed in Mandarin conversation (Tao 1996), which are unique characteristics in spoken language different from syntactic units. They are also found to be highly related to the language planning process. Moreover, it has been observed (Tseng 2006a) that almost all reparandum parts are themselves intonation units. The behavior of overt repair is just similar to that of a new intonation unit with respect to the preceding one. All these imply that overt repairs have a lot to do with the intonation units and thus pitch-related features. All these are consistent with the results here, i.e., the cues carried by pitch-related features provide important information for overt repair detection. On the other hand, we can also find that for direct repetition IP detection, the duration-related features are more important, and for abandoned utterances IP detection, both pitch-related and duration-related features have equally important impact.

8.2 Pitch-related features for IP detection of different disfluency types

The 13 features found to be the most important in IP detection for the four different types of disfluencies are represented by symbols (a) to (m) with their definitions as listed in Table 2, where the upper and lower halves are for pitch-related and duration-related features respectively. In Table 3, for each of the four categories of disfluencies, we list the symbols for the two pitch-related features causing the most serious recall rate degradation, or the two most important pitch-related features, together with the associated recall rate degradation, in the columns labeled as “pitch-related”. We can see that the average pitch value within a syllable, used in features represented by symbols (b) and (d) in Table 2, appears to be very important in three out of the four types of disfluencies,

regardless of different smoothing methods used. This suggests that the level of pitch is a very good cue for disfluency IP detection, probably due to the tone information carried and the intonation unit property as mentioned earlier. Especially, the absence of this feature degrades the performance very severely on partial repetitions and abandoned utterances.

Direct repetition, on the other hand, is much less influenced. Moreover, the difference of maximum and minimum pitch values within a syllable, used in features represented by (e) and (f) in Table 2, is beneficial to IP detection of direct repetitions and partial repetitions. It has been found (Tseng 2006a) that as far as Mandarin Chinese is concerned, the overt repairs, direct repetitions, and partial repetitions tend to be shorter. The main reason is probably that in Mandarin Chinese there is no inflection and the word order can vary to a great extent, speakers can re-initiate at the morphological boundary immediately after some inappropriateness is sensed. Moreover, it was also found that simple direct repetition repeating only one syllable usually dominates (Tseng 2006a). With plenty of such mono-syllable repeats, the pair of (partially) repeated and re-initiated syllables very often exhibit highly similar pitch contours. With the tone information inside these contours, pitch level (features (b) and (d)) and range (features (e) and (f)) can thus capture the evidence of short direct repetition and partial repetition. Another important pitch-related feature in Table 3 is the difference of pitch value across boundaries (used in the feature represented by (a)). This feature somehow conveys to what degree the speaker resets the pitch at this boundary. The reset of pitch is often the evidence of starting a new intonation unit, which is probably also the beginning of a new planning unit. This may be the reason why this feature is very important in the detection of abandoned utterances and overt repair IPs.

Table 2: The definitions of features used in Table 3. Δ(z): the parameter z was evaluated for each syllable boundary, and Δ(z) is the difference of the parameter z for two neighboring syllable boundaries.

Pitch-related features	(a)	Δ(difference of pitch slope across boundary)
	(b)	Δ(average pitch value within a syllable), with pitch value obtained from raw f0 value
	(c)	averaged absolute value of pitch slope within a syllable, with pitch value obtained from linear approximation
	(d)	Δ(average pitch within a syllable), with pitch value obtained from PCA
	(e)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from raw f0 value
	(f)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from linear approximation

Duration-related features	(g)	Δ(ratio of the duration for the syllable before the boundary to the pause duration at the boundary)
	(h)	ratio of the duration for the syllable after the boundary to the pause duration at the boundary
	(i)	product of the duration for the syllable after the boundary with the pause duration at the boundary
	(j)	Δ(product of the duration for the syllable after the boundary with the pause duration at the boundary)
	(k)	Δ(ratio of the duration for the syllable after the boundary to the pause duration at the boundary)
	(l)	syllable duration parameter ratio across the boundary, with the duratoin parameter being the average over 3 neighboring syllables
	(m)	standard deviation of (product of the duration for the syllable before the boundary with the pause duration at the boundary)

Table 3: The recall rate degradation when excluding an pitch-related/duration-related feature for different types of disfluencies (with definitions of features listed in Table 2).

Disfluency Types	Most Important Features (recall degradation)		Second Important Features (recall degradation)	
	pitch-related	duration-related	pitch-related	duration-related
abandoned utterances	(a) (-17.25)	(g) (-17.25)	(b) (-14.97)	(h) (-14.97)
overt repairs	(c) (-26.67)	(i) (-13.33)	(a) (-20.00)	(j) (-13.33)
direct repetition	(d) (-5.40)	(k) (-8.10)	(e) (-5.40)	(l) (-8.10)
partial repetition	(b) (-18.21)	(h) (-16.33)	(f) (-18.21)	(m) (-16.33)

8.3 Duration-related features for IP detection of different disfluency types

Table 3 also listed similar analysis with respect to duration-related features, in which we list the two most important duration-related features, together with the associated recall rate degradation, for the four types of disfluencies, in the columns labeled as "duration-related". Although duration-related features are beneficial to direct repetition detection as mentioned above, they also help indicate IP of other types of disfluencies. First, jointly considering both the syllable duration and pause duration was shown to be useful across all kinds of disfluencies. Combining through ratio of syllable duration to pause duration (represented by (g), (h) and (k) in Table 2) is relevant to IP detection of abandoned utterances, direct and partial repetitions, while overt repairs and partial

repetition benefit from the product of them (represented by (i), (j) and (m) in Table 2). The ratios may have normalized the syllable duration with respect to the breathing tempo of the speaker, if any, which was revealed by the pause duration fluctuation. The results showed that such features are actually useful. Moreover, a specific feature for direct repetition is the character duration ratio across boundary (represented by (l)), implying how the speaking rate was fluctuating. This showed that direct repetitions usually cause significant speaking rate deviation, and this is consistent with the observation obtained before (Tseng 2006a), in which it was concluded that the repeated words in the resumption are shorter than those in the reparandum part, because the direct repetition itself often provides no new information. Partial repetitions also exhibit similar properties to those of direct repetition. The contribution of standard deviation (represented by (m)) to partial repetition may thus be also due to the duration fluctuation related to partial repetition. Although the effect of standard deviation (feature represented by (m)) on direct repetition is not shown on Table 3, it indeed stands right behind (being the third important, not shown in the table), which supports the above argument.

9. Experimental results

9.1 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. Fig. 5a compares IP detection accuracies obtained using the delicate utterance-type LPM-based models as shown in Fig. 2 to those using plain integrated 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.

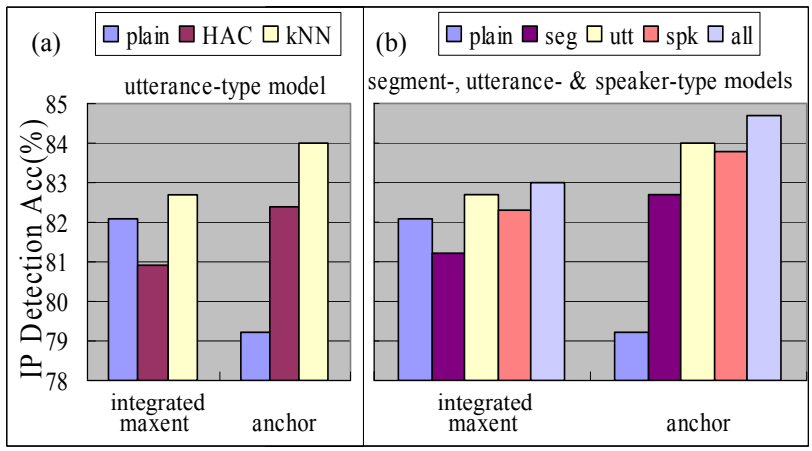

Figure 5: IP detection accuracy using LPM-based integrated maxent or anchor models, (a) with and without the utterance-type delicate models with training data selected by HAC- and kNN-based approaches. (b) with segment-, utterance-, and speaker-type delicate models with training data selected by kNN-based approach.

Fig. 5b 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 Fig. 5b for each case are the same as those in Fig. 5a. 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.

9.2 LPM-based feature expansion for integrated maxent

As mentioned in §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 integrated maxent model. The results for such case are shown in Fig. 6, where the bar (a) is the same as the last bar for integrated maxent in Fig. 5b, 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 integrated maxent model is combined with the anchor model by SVM, which eventually yielded the best result.

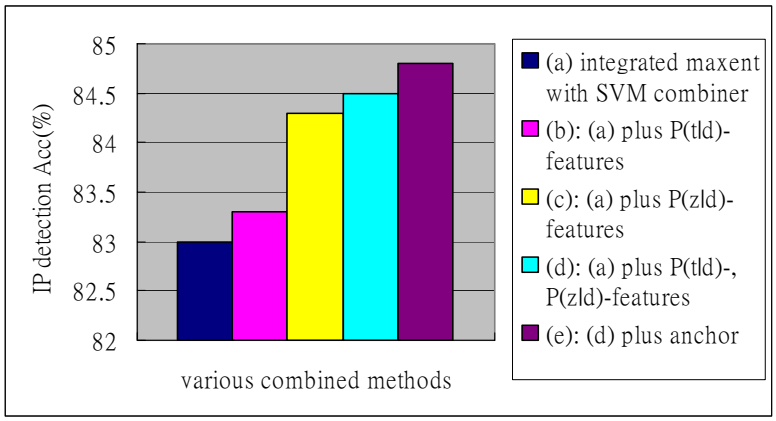

Figure 6: The performance of the integrated maxent model with expanded LPM-based features and finally integrated with the anchor model

9.3 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). Fig. 7 shows the character accuracy with IP detection results considered as a function of the weight parameters λ in equation (6), using the formula described in §7 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.

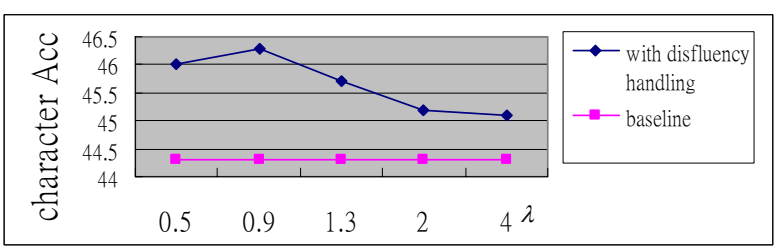

Figure 7: Character accuracy with disfluency IP detection.

10. Conclusions

We presented a new approach of modeling prosodic information in speech, LPM, for 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 integrated maxent and anchor models were enhanced by LPM. The results also verified the benefit of embedding disfluency information in the recognizer.

Che-Kuang Lin
Graduate Institute of Communication Engineering
National Taiwan University
Taipei 106, Taiwan
kimchy@speech.ee.ntu.edu.tw

Shu-Chuan Tseng
tsengsc@gate.sinica.edu.tw

Lin-Shan Lee
lslee@gate.sinica.edu.tw