Improved Spontaneous Mandarin Speech Recognition by Disfluency Interruption Point (IP) Detection Using Prosodic Features

Che-Kuang Lin and Lin-Shan Lee

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

Abstract
In this paper, a new approach for improved spontaneous Mandarin speech recognition with disfluencies well considered is presented. The basic idea is to detect the disfluency interruption points (IPs) prior to the recognition, and then to use these information during rescoring in the recognition process. For accurate detection of disfluency interruption points (IPs), a whole set of new features were proposed and tested by carefully considering the special characteristics of Mandarin Chinese. A new approach of incorporating the decision tree into the maximum entropy model training was also developed to enhance the IP detection accuracy. Experimental results indicated that the proposed set of features and the IP detection approach were very useful, and the obtained information about disfluency actually benefited the speech recognition performance.

1. Introduction
Most speech recognition systems can successfully process well-formed and well-spoken utterances. However, for ill-formed utterances frequently appearing in spontaneous conversation, properly modeling the ill-formness is a very important but still very difficult problem. One of the primary sources of ill-formness is the presence of disfluencies. Accurate identification of various types of disfluencies and properly utilizing such messages can not only improve the recognition performance, but provide structural information about the utterances.

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. The purpose of the research presented in this paper is to develop efficient approaches in accurate detection of such disfluency interruption point (IP) in spontaneous Mandarin speech, and use such information to improve the performance in speech recognition.

Consider the following example in Mandarin Chinese:

shi4 jin4kou3 EN chu1kou3 ma1?

(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.

2. Prosodic features
Due to the mono-syllabic structure of 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. Therefore, we 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 [6,7], and it has been found [4] 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 mono-syllabic structure as mentioned above and 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.

2.1. Pitch-related features
Pitch information is typically less robust and more difficult to use [6]. Pitch contour stylization method has thus been used, and smoothing out the “micro-intonation” and tracking errors was found helpful for English [6,7]. 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.

2.2. Duration-related features

Duration features such as pause and phone duration features have been used to describe prosodic continuity and preboundary lengthening [6,7]. 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.

3. Interruption point (IP) detection

The approaches used for IP detection are discussed in this section. The IP detection task is considered as a two-class classification problem here in this research. For each syllable boundary, a decision between “non-IP” vs. “IP” was made. Because IPs were relatively rare events, we used ensemble sampling [8] 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>train(7.1hr)</th>
<th>test(1.1hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of IPs / non-IPs</td>
<td>3569/92189</td>
<td>53674231</td>
</tr>
<tr>
<td>Chance of non-IPs</td>
<td>96.3%</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

Table 1: The summary of experiment data.

3.1. Decision tree (DT) and maximum entropy model (Maxent)

In the first approach, we used decision trees (DT) to learn from data, and to make prediction while testing [7]. 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 (Maxent) to make the decision [9]. 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:

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

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

The Maxent model is estimated by finding the parameters \( \lambda_i \) 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.

3.2. Integration of DT and Maxent

Considering the decision trees (DT) and maximum entropy model (Maxent) mentioned above, we can find that each of them has some advantages and limitations in dealing with the problem here. DT can handle real-valued features directly while Maxent is working in a discrete style. On the other hand, by carefully designing the feature function, Maxent may make finer decision on a certain feature parameter, while for DT 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 DT 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 DT and Maxent together. In this approach, we used decision trees built with training data to derive the feature functions for Maxent, 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 Maxent. 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 integrated DT and Maxent (DT-Maxent) in this paper.

4. Speech recognition with IP detection

Here we present the way to incorporate the output of the IP detection into the speech recognition processes. The IP
5. Experimental results

5.1. Corpus

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) [10], 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 groups of spontaneous speech phenomena. The 8 hrs of annotated dialogues as mentioned above were used in this research. Table 1 summarizes the data used in the following experiments. As can be found, 96.3% and 96.4% of the syllable boundaries are non-IPs.

5.2. IP detection using DT with feature analysis

To handle the data skewness issue intrinsic to IP detection problem, we created additional data set by downsampling the original imbalanced test data into balanced ones. The results of IP detection using DT with different feature sets are shown in Figure 2(a), where the results of both cases of assuming known transcription (ktr) and using recognized results with recognition errors (rec) are shown, with either the original imbalanced (Part A) or downsampled balanced data sets (Part B). Feature set 1 is exactly the same set used in previous work [6,7]. Feature set 2 is the same set but extracted for syllables rather than words as in set 1, to consider the monosyllabic structure of Mandarin Chinese, as compared to the feature sets proposed here in this paper. We see that feature set 2 performed much better than set 1 in all cases, or extracting features for syllables was better than words, apparently due to the monosyllabic structure of Mandarin Chinese. However, the feature set proposed here in this paper performed always the best. Also, by comparing the results of assuming known transcription (ktr) to that of using recognized results (rec), we can see the proposed feature set is the most robust with respect to recognition errors, while feature sets 1 and 2 are more sensitive to such errors. We also identified and investigated the contribution of the five most important features by examining the detection accuracy degradation when deleting one feature associated with the five top level nodes (those closest to the root) in the whole process as in Figure 2(b). For example, feature (i) is “product of the duration for the syllable after the boundary with the pause duration at the boundary”, and feature (ii) and (iii) are “difference of maximum and minimum pitch value within a syllable”, with pitch value obtained from raw f0 value and PCA respectively. The definition of feature (iv) to (v) is left out due to space limitation. As can be found, deleting feature (i) or (ii) degraded the accuracy from 73.3% to 55.2% and 61.8% respectively. They are certainly the key features. Note that four out of the five most important features are the new features proposed here in this paper. For further analysis of the proposed features, please refer to [11].
5.3. Comparison of different IP detection approaches

To compare different IP detection approaches, the detection accuracy for the three different approaches presented in Section 3, i.e., the decision tree (DT), maximum entropy modeling (Maxent) and integrated DT and Maxent (DT-Maxent), all using the new feature sets proposed here in this paper, are shown in Figure 3(a), where the notations ktr, rec and Part A are in parallel with those in Figure 2(a), except here the comparison is among IP detection approaches. Apparently Maxent is better than DT, but the proposed DT-Maxent approach offered much better detection accuracy. For all the above results, we trained the models via balanced training data, which implied severe mismatch between training and testing situations. In some earlier work [7], it was proposed to use prior information of the training data to combat this problem. Instead of using the true prior distribution ratio in the training data of 96:4, we chose to use a prior ratio of 75:25 to moderately help with this problem (i.e. (probability of non-IP)*0.75+(probability of IP)*0.25 and so on) without overwhelming the decision made by models. The results are in Figure 3(b) of Part C, where we can see the real strength of the proposed features and approaches.

5.4. Speech recognition results

The recognition experiments were performed with a lexicon of 50K entries, a trigram language model, and an intra-syllabic right context dependent Initial/Final acoustic model set (a Mandarin syllable was decomposed into two parts: Initial and Final). Figure 4 shows the character accuracy with IP detection results considered as a function of the weight used for the IP probability in the rescoring process, compared to the baseline without considering the disfluency. We see that the highest improvement can be achieved at the IP probability weight being about 1.3. Table 2 shows the substitution, deletion, and insertion rates for characters. One-pass search in row (a) the baseline, we see the disfluency detection(row (b)) actually reduced the substitution error rate. We also observed the relatively high insertion error rate in both rows (a) and (b), apparently caused by the very frequently appearing non-speech sound such as coughs and laughter. The results for trigram resoring only in row (c) serving as a reference, by comparing with row (b), we see rather significant improvement was actually obtained by the information of disfluency in the rescored recognition process.

6. Conclusions

We presented a new approach of spontaneous Mandarin speech recognition by carefully detecting the disfluency IP. The disfluency IP detection used a set of new prosodic features considering the characteristics of Chinese language, and the best IP detection was achieved by incorporating decision trees into the maximum entropy model training. The IP information was then used in rescoring in the recognition process. The experiments verified the benefit of embedding disfluency information in the recognizer.

7. Acknowledgements

The authors wish to thank Dr. Shu-Chuan Tseng for providing the MCDC corpus for experiments and the very helpful discussions.

8. References