*Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 117-121.*

Important and New Features with Analysis for Disfluency Interruption Point (IP) Detection in Spontaneous Mandarin Speech

Che-Kuang Lin, Shu-Chuan Tseng, Lin-Shan Lee*

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan
Institute of Linguistics, Academia Sinica, Taipei, Taiwan*

Abstract

This paper presents a whole set of new features, some duration-related and some pitch-related, to be used in disfluency interruption point (IP) detection for spontaneous Mandarin speech, considering the special linguistic characteristics of Mandarin Chinese. Decision tree is incorporated into the maximum entropy model to perform the IP detection. By examining performance degradation when each specific feature was missing from the whole set, the most important features for IP detection for each disfluency type were analyzed in detail. The experiments were conducted on the Mandarin Conversational Dialogue Corpus (MCDC) developed by the Institute of Linguistics of Academia Sinica in Taiwan.

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 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. The disfluencies considered here in this paper include the following four categories:

(1) 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.

(2) Partial repetitions: only part of a word (including compound words) is repeated.

(3) Overt repairs: the speaker modifies expressed words within utterance.

(4) Abandoned utterances: a speaker abandons an utterance and starts over.

Consider the following example (overt repairs):

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 analyzed here.

Consider another example (direct repetition):

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.

It has been suggested much earlier [2] that there exists a certain acoustic "edit signal" serving as a cue indicating that fluent speech had been interrupted. Although it may be difficult to find a single cue for such purposes, several prior studies had indicated that combinations of more cues can be used to identify disfluencies with reasonable success [3,9,10].

In this paper, a whole set of acoustic-prosodic features were considered and disfluency interruption point (IP) detection was tested and analyzed. Two types of features were considered here, duration-related and pitch-related. Detailed analysis regarding which features are the most important for which types of disfluencies and possible reasons are then discussed. Below the corpus used in this research is first introduced in section 2, while the acoustic-prosodic features investigated are summarized in section 3. Section 4 briefly describes the two approaches used for the disfluency interruption point (IP) detection. Section 5 finally presents the experimental results and relevant analysis.

2. Corpus Used in the Research

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. Several studies have been conducted on MCDC to analyze various phenomena of spontaneous conversational Mandarin speech [13,14,15]. 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: disfluency, 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 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. As can be found, 96.3% and 96.4% of the syllable boundaries are non-IPs. The total number of IPs is limited in the annotated corpus, which makes the studies and analysis slightly difficult.

Table 1: The summary of experiment data.

	train	test
Data length	7.1hr	1.1hr
Number of non-IPs	92189	14231
Number of IPs	3569	536
Chance of non-IPs	96.3%	96.4%

3. Prosodic Features

We tried to define a whole set of acoustic-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 [5,11], 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 were 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 [11], partly due to the variabilities of pitch values across speakers and speaking contexts, partly due to serious pitch tracking errors. A pitch contour stylization method has thus been used and smoothing out the “micro-intonation” and tracking errors has been found helpful for English [5,11]. 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 still certainly needed. Syllable-wise pitch contour smoothing and Principal Component Analysis (PCA) have both been shown to be helpful in identifying key characteristics in the pitch contours and performing the tone recognition in Mandarin Chinese [8,12].

We used 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. Several syllable-wise pitch-related features found useful in tone recognition [8] 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.

3.2. Duration-related Features

Duration features such as pause and phone duration features have been used to describe prosodic continuity and preboundary lengthening [5,11]. By carefully examining the characteristics of IPs in our corpus, we hypothesized that deviation from 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 fluctuation 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. Other groups 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.

4. IP Detection

We use the acoustic-prosodic features mentioned above to detect IP events given all syllable boundaries after the first pass recognition giving all the recognized syllables. Two analytical approaches were used in the IP detection, the decision tree and the maximum entropy model. The task here is simplified as a two-class classification problem. For each syllable boundary, a decision of “non-IP” vs. “IP” is made. Due to the very limited number of IPs of different disfluency types in the available annotated corpus, they were grouped together as a single class of “IP”, and all other boundaries then belong to the class of “non-IP”. Because IPs are relatively rare events, the approach of ensemble sampling previously proposed [6] was used on the training data to equate the prior probabilities for the two different classes. This made the approaches more sensitive to any inherent prosodic features that can distinguish the classes. In the first set of experiments, we used decision trees [5] to learn from the data, to identify the useful features, decide how to use them and finally use them to detect the IPs. The decision was made based on the posterior probabilities for the leave nodes where the syllable boundary being considered went to. In the second approach of the maximum entropy model [1], on the other hand, each feature is expressed by a binary feature function, and the model tried to find the appropriate parameters for each feature function with the constraint that the expected values of the various feature functions match the empirical averages in the training data. Some improved approaches were then developed trying to integrate the nice properties of the two different approaches, in which the threshold values for the various features learned from the decision tree were used as the quantization thresholds in defining the feature functions used in the maximum entropy model. This approach turned out to be better than the original maximum entropy model. The results reported below for maximum entropy model were obtained by this integrated approach.

5. Experiment Results & Further Analysis

5.1. IP Detection Results

The IP detection results in terms of recall and precision rates using decision tree and maximum entropy model are listed in Table 2. While decision tree achieves moderate and balanced recall and precision rates, maximum entropy model trades degraded recall for significantly better precision. As far as performance of a recognition system is concerned, a false alarm is usually more harmful than an omission. In other words, for the purposes here it is preferred to achieve as high precision as possible while having a high enough recall rate. As a result, maximum entropy model may be more appropriate for the purpose here.

Table 2: Recall and precision rates for IP detection with decision tree and maximum entropy model.

	Recall	**Precision**
Decision Tree	73.15	73.03
Maximum Entropy	56.38	81.95

5.2. 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, precision for each disfluency type was not obtainable, while recall was. Only the results of maximum entropy model were discussed here due to space limitation, although almost the same trend was found in the decision tree approach.

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 Figure 1, 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 find that for overt repair and partial repetition, pitch-related features play relatively more important role for IP detection, and this is especially apparent for overt repair. This is in good consistency with the earlier findings [16] 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 [17], 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 [16] 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 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.

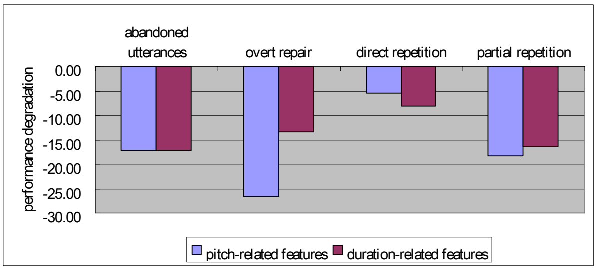

Figure 1: Performance degradation (recall degradation) for the four disfluency types with respect to the two feature categories.

5.3. 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 3, where the upper and lower halves are for pitch-related and duration-related features respectively.

In Table 4, 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 3, 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. In particular, 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 3, is beneficial to IP detection of direct repetitions and partial repetitions.

It has been found [16] 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 [16]. With many 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 4 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 3: The definitions of features used in Table 4. Note that for certain parameter z evaluated for each syllable boundary, Δ(z) is the difference of the parameter z for two neighboring syllable boundaries.

	Feature ID	Definition
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 duration 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 4: The recall rate degradation when excluding an pitch-related/duration-related feature for different types of disfluencies. (with definitions of features listed in Table 3).

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)

5.4. Duration-related Features for IP Detection of Different Disfluency Types

Table 4 also showed 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 3) 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 3). 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 [16], 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 4, it indeed stands right behind (being the third important, not shown in the table), which supports the above argument.

6. Discussion

A whole set of features to be used for disfluency IP detection is developed, tested and analyzed. The most important features for each disfluency types were identified and discussed considering the linguistic characteristics of the disfluencies. The false alarms obtained in the detection output remains to be a major problem for further applications. One possible approach toward this direction is probably to adopt some kind of disfluency type-specific rule-based methods to further discriminate the false alarms.

7. References

[1] Berger, A. L., Della Pietra, S. A., & Della Pietra, V. J. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39-72.

[2] Hindle, D. 1983. Deterministic Parsing of Syntactic Nonfluencies. *Proc. ACL'83*, pp.123-128.

[3] Lickley, R. J. 1996. Juncture Cues to Disfluency. *Proc. ICSLP'96.*

[4] Liu, Y., et al. 2005. Structural Metadata Research in the EARS Program. *Invited paper. Proc.ICASSP'05.*

[5] Liu, Y., Shriberg, E., & Stolcke, A. 2003. Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources. *Proc. Eurospeech'03*, pp. 957-960.

[6] Liu, Y., Shriberg, E., Stolcke, A., & Harper, M. 2004. Using Machine Learning to Cope with Imbalanced

Classes in Natural Speech: Evidence From Sentence Boundary and Disfluency Detection. *Proc. of ICSLP'04.*

[7] Lin, C.-K. & Lee, L.-S. Improved Spontaneous Mandarin Speech Recognition by Disfluency Interruption Point (IP) Detection Using Prosodic Features. *Proc. Eurospeech'05.* (to appear).

[8] Lin, W.-Y. & Lee, L.-S. Improved Tone Recognition for Fluent Mandarin Speech Based on New Inter-syllabic Features and Robust Pitch Extraction, *Proc. ASRU'03*.

[9] Nakatani, C. & Hirschberg, J. 1994. A Corpus-based Study of Repair Cues in Spontaneous Speech. *JASA*, pp.1603-1616, 1994.

[10] Shriberg, E. 1999. Phonetic Consequences of Speech Disfluency. *Proc. ICPhS'99*, pp. 619-622.

[11] Shriberg, E., et al. 2000. Prosody-based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, pp. 127-154, 2000.

[12] Tian, J. & Nurminen, J. 2004. On Analysis of Eigenpitch in Mandarin Chinese. *Proc. ISCSLP'04*.

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

[14] Tseng, S.-C. 2003. Repairs and Repetitions in Spontaneous Mandarin. In *Proceedings of Workshop on Disfluency in Spontaneous Speech (DISS 03)*. Ed. Robert Eklund. Gothenburg Papers in Theoretical Linguistics 90. pp. 71-74. University of Gothenburg.

[15] Tseng, S.-C. 2005. Syllable Contractions in a Mandarin Conversational Dialogue Corpus. *International Journal of Corpus Linguistics*. 10(1): 63-83.

[16] Tseng, S.-C. Repairs in Mandarin Conversation. *Journal of Chinese Linguistics.* (to appear).

[17] Tao, H.-Y. 1996. *Units in Mandarin Conversation*. Prosody, Discourse, and Grammar. John Benjamins Publishing Company.