Detection of Word Fragments in Mandarin Telephone Conversation

Cheng-Tao Chu[1], Yun-Hsuan Sung[2], Yuan Zhao[3], Dan Jurafsky[3]

Departments of Computer Science[1], Electrical Engineering[2], Department of Linguistics[3]
Stanford University, Stanford, CA 94305

chengtao@cs.stanford.edu;{yhsung,yuanzhao,jurafsky}@stanford.edu

Abstract

We describe preliminary work on the detection of word fragments in Mandarin conversational telephone speech. We extracted prosodic, voice quality, and lexical features, and trained Decision Tree and SVM classifiers. Previous research shows that glottalization features are instrumental in English fragment detection. However, we show that Mandarin fragments are quite different than English; 90% of Mandarin fragments are followed immediately by a repetition of the fragmentary word. These *repetition fragments* are not glottalized, and they have a very specific distribution; the 12 most frequent words ("you", "I", "that", "have", "then", etc.) cover 50% of the tokens of these fragments. Thus rather than glottalization, we found the most useful feature for Mandarin fragment detection was the identity of the neighboring character (word or morpheme). In an oracle experiment using the true (reference) neighboring words as well as prosodic and voice quality features, we achieved 80% accuracy in Mandarin fragment detection.

Index terms: Disfluencies, Word Fragments, Mandarin, Prosody, Voice Quality

1. Introduction

Fragments are parts of words that result from the speaker breaking off in the middle of word production. Although the raw frequency of fragments isn't extremely high (just under 1 per 100 words in the Switchboard corpus [1]) they are good indicators for other kinds of disfluencies, such as fillers and restarts. In the ATIS corpus, for example, Bear et al [12] and Nakatani and Hirschberg [5] found that 60-74% of disfluencies contained a word fragment. Fragment detection can thus play an important role in improving disfluency annotation and perhaps also word error improvements. But since any word can be produced as a fragment, and since word fragments are quite short, fragments are difficult to model in the standard HMM lexicon, and hence are generally misrecognized in LVCSR systems.

Natakatani and Hischberg [5] analyzed repairs cues in English speech based on acoustic and prosodic cues and proposed a number of features that could be used for fragment detection. Following this line of work, Liu [1] showed that fragments in English could be detected with 72.9% accuracy using only prosodic and voice-quality features. Many of the prosodic features she used were drawn from Shriberg et al [2], who showed their usefulness in speech tasks including sentence boundary detection, disfluency detection, and topic segmentation.

Mandarin has word fragments as well, and an initial investigation using the SONIC LVCSR system [6] described below suggested that they are generally misrecognized, either via substitution or deletion, as shown in the following examples:

Substitution:	你 - 你 下次	跟他 说
	you-you next time	to him tell
Recognizer output:	那 你 下次	跟他 说
	that you next time	to him tell
Deletion:	对 如 - 如果 哦	
	Right ***if*** -if	oh
Recognizer output:	对 如果	哦
	Right if	oh

In the substitution example, the character 你 *ni* 'you is misrecognized as 那 *na* 'that', while the character 如 *ru* 'if' is deleted in the second example.

In this paper we propose to extend the insights of Liu [1] and Nakatani and Hirschberg [5] and build classifiers to distinguish between fragments and non-fragments in Mandarin. The next section introduces our corpus. Section 3 describes our first experiment using features derived from the English work. Error analysis of experiment results and suggestions are reported in Section 4. In Section 5 we extract some new features suggested by the error analysis; we finish with some conclusions.

2. Corpus

We used the Hong Kong University of Science and Technology (HKUST) 200-hr Mandarin Chinese conversational telephone speech corpus [8]. Speakers were chosen from several cities across Mainland China and include standard and accented speakers. There are 874 and 25 conversations in the training and development datasets, respectively. Most conversations are 10 minute long. The training data conversations sides include 949 male speakers and 797 female speakers. There are in total 1,455,030 characters and 116,590 sentences; 12.48 characters per sentence. The HKUST transcriptions mark a word with a dash when "the speaker breaks off in the middle of the word" [8]. We thus classified as fragments any word that ended in a dash in the transcript. There were 13820 fragment characters in the 1,455,030 total characters, i.e., 9.5 fragments per 1000 characters.

3. Experiment 1: English Features from Liu [1]

Beginning with the hypothesis that Mandarin fragments behave like English fragments, in our first experiment we propose to replicate Liu's [1] work, but on the detection of Mandarin rather than English fragments.

3.1 Experiment Setup

Following the paradigm of Liu [1], we built classifiers which, given a boundary between words, determined whether the word preceding the boundary was a fragment or not. We extracted features for the words before and after the boundaries.

In fact, we actually considered both words and characters as the basic unit of segmentation in all the experiments in this paper, and so each experiment is reported with features extracted across both words and characters. Because the HKUST corpus is not word-segmented, we used the maximal matching algorithm to segment each sentence into words.

In each of our experiments, we first force-aligned the reference (correct) transcripts to the speech waveform, using the Sonic LVCSR system [6]. We chose to use this kind of oracle information, following Liu [1], because we felt that the current WER on Chinese continuous speech was too low to get useful word identities and boundaries. We next extracted features at each boundary, from both the previous and following character or word segment. The fragment word segment is the time region from current boundary to previous one. For example, the word segment of 我 C_1 is from B_0 to B_1.

B_0 B_1 B_2 B_3 B_4 B_5 B_6 B_7 B_8 B_9 B_{10}
| C_1 | C_2 | C_3 | C_4 | C_5 | C_6 | C_7 | C_8 | C_9 | C_{10} |
我 - 我 问 的 是 有 什么 影 响
I - I ask DE copula have what influence
'I-I asked what influence it has.'

Because the number of non-fragments is vastly higher than the number of fragments, we randomly downsampled the whole training set to 1000/1000 fragment/non-fragment data as training data and 100/100 fragment/non-fragment data as testing data. We used both decision tree (C4.5) and SVM (LIBSVM [4]) classifiers.

In the SVM, before training, we first scale all features into the region [0,1]. This step prevents any feature from being weighted more heavily than others (since we assume that each feature is equally important). During training, we use the RBF kernel; parameters are grid searched automatically. The best parameters are chosen based on 5 fold cross validation. After obtaining the parameters we scale the test data by the same amount as the training data.

3.2 Features for Expt 1: Prosody + Voice Quality

Our initial experiment aims at replicating features that prove to be useful in English fragment detection. We employed the prosodic features suggested by Liu [1], as well as voice quality features that she designed to detect glottalization (since English fragments are known to be glottalized). We extracted each feature in two ways, over a character region and over a word region:

Prosodic features: pitch, energy, and duration.

We extract the **pitch** contour in each word/character via Praat [7] and use statistics such as the mean, minimum, maximum and slope as features. Because fragments are partial words, we expect there are obvious changes across fragment boundary. We stylized pitch contour, using a simplified version of [7] [10] and extract the slope difference between the pre-fragment segment and post-fragment segment.

For **intensity** features, we first extract a signal using a Hamming widow. Then we calculate the energy of that windowed signal and stylize the energy contour. We use the same features for energy as for pitch.

For **duration** features, we used word normalized durations and pulse durations computed from the forced alignment.

Voice quality features, all directly following Liu [1]

Open quotient (**OQ**) is defined as the ratio of the time when the vocal folds are open to the total length of a glottal cycle, which is used to detect modal voicing, creaking voicing, and breathy voicing. Liu [1] found that fragments seemed to exhibit OQ indicative of creaky and breathy voice. Fant [3] formulated open quotient by regression analysis as:

$$OQ = \log((H_1^* - H_2^* + 6)/0.27)/5.5$$

Where H_1^* and H_2^* are the spectral intensities, in decibels, of F0 and twice-F0. We used average, minimum, and maximum within a windowed region as features.

Spectral Tilt is the slope of the spectrum of the speech signal, which can also help model phonation type. We extracted formants and intensity at each formant frequency. Spectral tilt is calculated via linear regression. We also used average, minimum and maximum as features.

Jitter measures the irregularity of pulses around the boundary. Our measurement uses Praat's ddp method [7]. We calculated the average absolute difference between consecutive differences between consecutive periods, divided by the average period, as follows:

$$jitter = \frac{\sum_{i=2}^{N-1} |2 * T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i}$$

3.3 Results of Experiment 1

Table 1 and Table 2 show that using both prosodic features and voice quality measures we achieve 65.5% and 69.5% in word-based and character-based SVM, respectively, suggesting three conclusions. First, extracting features over neighboring characters rather than entire words improves classification (by 7.5% and 4.0 % absolute in DT and SVM, respectively). Therefore, in the following experiments, we only use characters as basic unit. Second, our results using DT are roughly 6% worse than Liu's [1] comparable DT fragment detection results in English. Third, we found open quotient and jitter less useful for Chinese than Liu found for English in [1].

Table 1. *Word level results.*

Word-level	DT	SVM
Prosodic + Voice Quality	59.50 %	65.50 %
Prosodic only	59.00 %	64.50 %
Voice Quality only	59.50 %	63.00 %

Table 2. *Character level results.*

Character-level	DT	SVM
Prosodic + Voice Quality	67.00 %	69.50 %
Prosodic only	65.00 %	69.50 %
Voice Quality only	60.50 %	56.50 %

4. Error analysis

Why were our fragment detection rates lower in Chinese than Liu found for English, and why were open quotient and jitter not useful? Part of the difference between the English and

Chinese error rates is likely due to our simplified version of the pitch stylization of [7] [10]. But we also suspected that Mandarin fragments might differ from English ones in terms of their distribution, lexical complexity and acoustic properties. Thus in order to find robust cues for Mandarin fragments detection, we carried out a linguistic analysis of Mandarin repairs, examining a wide variety of the fragments from the corpus.

We found that the fragments could be largely categorized into two basic types: **lexical repetition fragments** (a) and **lexical alternation fragments** (b), as shown below:

a. 我 - 我 问 的 是 有 什么 影响
I - I ask DE copula have what influence
'I-I asked what influence it has.'
b. 他 却 很 - 活的 很好
he but very- live very well
'But he lives very well.'

Example (a) is a case of lexical repetition, since the repairing part is the same as the reparandum; (b) differs from (a) in that the lexical item in the repair differs from the reparandum.

We examined the distribution of lexical repetition fragments and lexical alternation fragments in both Mandarin and English corpora (using the Mississippi State transcripts of the Switchboard corpus [11]), as shown in Table 3:

Table 3: *Distribution of repetition and alternation fragments in English and Mandarin corpora*

	Repetition	Alternation
English	3241 (31.5%)	7044 (68.5%)
Mandarin	12522 (90.6%)	1298 (9.4%)

We found that in English the majority of repairs are lexical alternations (68.5%); Mandarin fragments, by contrast, are almost all lexical repetitions (90.6%); alternations only account for 9.4%. The distribution of repetition fragments was also quite skewed; the most frequent 12 characters account for 50% of the repetition fragments. The five most frequent fragments are summarized in Table 4.

Table 4: *Top five most frequent fragmented characters*

character	counts	Percentage
你 'you'	1651	11.96%
我 'I'	1604	11.62%
那 'that'	1022	7.40%
有 'have'	460	3.33%
就 'then'	334	2.41%

This result suggests that character identity in fragments is highly predictable in Mandarin. Adding character-level information thus may help improve the performance of the classifier.

To further investigate the differences, we checked the acoustic properties of Mandarin repairs. Previous literature reports that the most indicative acoustic features for English fragment detection are jitter and OQ [1]. Other features such as syllable duration are not queried much by the decision tree [1]. However our direct application of jitter and OQ in Mandarin fragment detection produced a result barely above chance, which might be attributed to two reasons. The first possibility is that our jitter and OQ based on forced alignment results might not be reliable, since some boundaries are not accurately segmented. Another possibility is that glottalization around the cut-off points may not be the most indicative cue for Mandarin fragments. Instead other prosodic features such as duration might be more salient.

In order to test the hypothesis, we took a random sample of 377 repetitions and 339 alternations. We automatically extracted the syllable duration before the pause (reparandum) and the syllable duration after the pause (repair) from the forced alignment results. In addition, we calculated the jitter (ddp) and OQ of the interruption point from a larger set of repetitions and alternations data. The average values are shown in Table 5:

Table 5: *Average jitter, OQ, duration pre/post-pause*

		Pre-pause syllable	Post-pause syllable
Repetition	Dur	307	262
	Jitter	0.0397	
	OQ	0.3732	
Alternation	Dur	287	243
	Jitter	0.0462	
	OQ	0.4023	

Table 6: *Average jitter for fragments vs. non-fragments*

	Jitter
Fragments	0.042
Non-fragments	0.034

We found that the syllable duration of the reparandum is significantly longer that that of the repairing part ($t_{repetition}(376)=36.073$, $p<.000$); $t_{alternation}$ $(338)=31.236$, $p<.000$). The average jitter values for fragments and non-fragments as shown in Table 6, however, do not show much difference, suggesting that voice quality features may not be very useful in differentiating fragments from non-fragments in general. Instead, prosodic features such as duration might be a better cue.

In summary, our error analysis suggests building separate repetition and alternation classifiers, and indicates lexical features should help in fragment detection, while voice quality features may not help for Chinese although they did for English.

5. Improved experiment

5.1 Lexical Features

As shown Table 3, most fragments are repetitions, suggesting that word identity may help in fragment detection. We also found that most of the repeated words are highly predictable frequent words (often pronouns). Therefore we incorporate lexical information on both word-level and character-level into our classifier. Instead of using all word identity as features, we only use the 100 most frequent word binary features. We added binary features for the presence of these words in the word/character before the fragment and the word/character after the fragment, totally 200 features. For example:

C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9 C_{10}
我 - 我 问 的 是 有 什么 影 响

For word C_2, we add word C_1 and word C_3 as features. If they are in the top 100 words, then the corresponding feature is set to one. Otherwise, all other binary word context features are zero.

The results are shown in Table 7. After adding context information into classifiers, we achieve 77.8% (DT) and 78.5% (SVM) using the character level, a 9~10% absolute improvement in both DT and SVM. In general, oracle lexical information helps greatly in fragment detection.

Table 7 *Character level results*.

	DT	SVM
All features	77.80%	78.50%
Prosodic + Voice Quality	65.80%	66.75%
Prosodic + Lexical	78.00%	80.00%
Voice Quality + Lexical	74.80%	78.75%
Prosodic Only	67.00%	69.25%
Voice Quality Only	56.20%	57.75%
Lexical Only	78.50%	78.25%

5.2 Two Classifiers

In section 4, we found that in Mandarin glottalization is not as useful a feature as for English fragments. In order to further investigate whether voice quality would help in the detection of any of the subtype of fragments, we separate our corpus into the two fragment types and use 1000/1000 and 600/600 observations to train repetition classifier and alternation classifier separately, respectively. Results are summarized in Tables 8.

Table 8. *Repetition and alternation fragment classifier*.

Repetition classifier	Repetition	Alternation
All features	79.50%	72.00%
Prosodic + Voice Quality	66.00%	76.00%
Prosodic + Lexical	81.50%	74.00%
Voice Quality + Lexical	79.00%	70.00%
Prosodic Only	65.50%	76.00%
Voice Quality Only	60.00%	62.00%
Lexical Only	81.00%	70.00%

We found that voice quality features do not help much in either of the two cases, suggesting that glottalization features in general do not help much in Mandarin fragment detection. The results are in accordance with our error analysis, which found very similar jitter values for fragments and non-fragments. The fact that Tone 3 and 4 syllables are known to be glottalized [13] might also degrade the performance of the voice quality feature. In addition, we found that lexical features significantly improve the classifier's performance in repetition detection, but not so much in alternations, since character identity of repetitions are more predictable than that of alternations. Finally our results show that prosodic features are more useful in detecting alternation fragments than repetitions.

6. Conclusion

In this paper, we built automatic detectors for word fragments in Mandarin conversational telephone speech, using prosodic, voice quality, and oracle lexical features, as well as alignments from oracle transcripts. We found that, like English, prosodic features are helpful in detecting fragments. Unlike English, voice quality features do not help much in Mandarin fragment detection. We suggest this is because Mandarin fragments are quite different than English; 90% of Mandarin fragments are followed immediately by a repetition of the fragmentary word. These *repetition fragments* are not much glottalized, and the repetitions have a very specific distribution; the 12 most frequent words ("you", "I", "that", "have", "then", etc.) cover 50% of the tokens of these fragments. Thus rather than glottalization, we found the most useful feature for Mandarin fragment detection was the identity of the neighboring character (word or morpheme). By adding (oracle) previous word and next word as lexical features into classifiers, we achieve 80.00% fragment detection accuracy.

Our results are only oracle results; our current goal is thus both to find new features to increase our performance on oracle data as well as move toward fragment detection from recognition output.

Acknowledgements

Many thanks to two reviewers and also to Yang Liu for advice, for generously answering many questions and making her code available. This work was partially supported by the Edinburgh-Stanford LINK and the ONR (MURI award N000140510388).

References

[1] Y. Liu, 2003. Word fragment identification using acoustic-prosodic features in conversational speech. HLT-NAACL student research workshop, 37-42.

[2] E. Shriberg and A. Stolcke. Prosody Modeling for Automatic Speech Recognition and Understanding. Mathematical Foundations of Speech and Language Processing.

[3] G. Fant. 1997. *The voice source in connected speech*. Speech Communication, 22:125—139.

[4] C-C. Chang and C-J. Lin. 2001. LIBSVM : a library for support vector machines. From http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5]C. Nakatani and J. Hirschberg. 1994. A corpus-based study of repair cues in spontaneous speech. JASA 1603—1616.

[6]B. Pellom, K. Hacioglu. 2003. Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", in Proceedings of IEEE ICASSP 2003, Hong Kong.

[7]P. Boersma & D. Weenik (2006): Praat: doing phonetics by computer (Version 4.4.16) [Computer program]. Retrieved from http://www.praat.org/ on Jan 22, 2006.

[8] P. Fung, S. Huang, D. Graff, HKUST Mandarin Telephone Transcripts, Part 1, LDC catalog number LDC2005T32.

[9] Transcription guidelines for EARS Chinese telephony conversational speech database.

[10] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. Proc. ICSLP, vol. 7, pp. 3189-3192, Sydney, 1998.

[11] http://www.isip.msstate.edu/projects/switchboard/releases/switchboard_word_alignments.tar.gz

[12] Bear, J., Dowding, J. & Shriberg, E.E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. *Proc. ACL*, pp. 56-63.

[13] Belotel-Grenie, A., & Grenie, M. (1994). Phonation types analysis in Standard Chinese. *Proc. ICSLP'94*, Yokohama, Japan, 343-346.