USING ACOUSTIC AND PROSODIC CUES TO CORRECT CHINESE SPEECH REPAIRS

Yue-Shi Lee and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, Republic of China
E-mail: {leeys, hh_chen}@csie.ntu.edu.tw

ABSTRACT

Speech repairs introduce much noise in spoken language processing. Properly correcting speech repairs can help the speech recognizer to avoid the textual errors, and prevent the interpretation errors during the subsequent processing. Because the task of repair processing cannot defer to the latter (word segmentation, part-of-speech tagging and sentence parsing) stages, this paper employs acoustic and prosodic cues to correct Chinese repetition and addition repairs. The experimental results show that the precision rate of 93.87% (76.09%) and the recall rate of 90.65% (70%) can be achieved for correcting Chinese repetition (addition) repairs.

1. INTRODUCTION

Repair is a very common phenomenon in spoken languages. Speakers usually repeat, add, replace, or even abandon some constructions in utterances for some mental reasons. A repair is composed of a repaired segment and a repairing segment which immediately follows the repaired segment. A repaired segment denotes the portion that the speaker utters need to be removed in order to correctly understand the speaker's meaning.

Heeman and Allen [1] describe that 25% of turns contain at least one speech repair in their corpus. In our study [2], 17% of turns contain at least one speech repair. Thus, the repair processing cannot be negligible and has influences to a certain extent.

On the one hand, correctly recognizing speech repairs can help automatic speech recognizers to avoid textual errors. In most of the current speech recognition systems, words repeated in a speech repair are simply fed as word hypotheses to the language model of the recognizer. This may cause difficulties in having a proper recognition since the language model is usually trained on fluent text only. On the other hand, even if all the words in a disfluent segment are correctly recognized, failure to detect a disfluency may lead to interpretation errors during subsequent processing. The details can refer to [2-3].

Recently, text-first approach and speech-first approach have been proposed to touch on repairs in English. The text-first approach assumes the speech recognizer could provide a correct transcription and tries to detect and correct speech repairs automatically using text alone. Bear, *et al.* [4] firstly try to parse the input sentence and then invoke a repair processing when the parsing fails. For repair processing, a simple pattern matcher finds the candidates based on the lexical cues at the first stage. Then the syntactic and semantic processing filters out the impossible candidates. Heeman and Allen [1] use repair patterns to capture potential repairs. These patterns are built based on the identification of word fragments, editing terms[1], and word correspondences between the repaired segment and the repairing segment. The resulting potential repairs are then passed to a statistical filter that judges the proposal as either fluent speech or an actual repair. The speech-first approach tries to identify repairs using acoustic and prosodic cues. Nakatani and Hirschberg [3] investigate the detection of the interruption point of speech repairs based on this line. The cues that they found are the occurrence of a filled pause, the presence of a word fragment, the energy peaks in each word and other features such as accent. However, they do not address the problem of correcting speech repairs. In other words, they do not determine which words are undesired.

These approaches cannot be adopted to deal with Chinese speech repairs for the following reasons. First, a Chinese sentence is composed of a string of characters without any word boundaries. That is, it is necessary to segment Chinese sentence before tagging and parsing. Repairs make segmentation and text-first approach more difficult. Second, Chinese repairs may not always have an editing terms between a repaired segment and a repairing segment. In other words, editing terms do not have much effect in Chinese repair processing. Third, duplicate constructions generated by repeating words or phrases in Chinese utterances are used very often, but they do not always initiate a repair. Forth, the Chinese speech repairs may be initiated at various syntactic environment [5], e.g., before the subject, during the subject, after the subject and before the verb, during the

[1] The editing terms can either be filled pauses (e.g., um, un, er) or cue phrases (e.g., I mean, I guess, well, let's see).

verb, and so on. The variety makes the identification of Chinese speech repairs more troublesome.

Because the repairs introduce much noise in language processing, we cannot defer the task of repair processing to the latter (word segmentation, part-of-speech tagging and sentence parsing) stages. Our previous work [6] has compared the performances of Chinese homophone disambiguation with and without repair processing. The experimental results show the significant performance improvements after treatments of Chinese repetition repairs. This paper will employ acoustic and prosodic cues to correct repetition and addition Chinese speech repairs. The latter is more complex than the former. New cues will be proposed in this paper.

2. TYPES OF CHINESE SPEECH REPAIRS

The speech repairs tend to have a standard form. There are four types of speech repairs in Chinese, i.e., repetition, addition, replacement and abandon. Let A, B, C and D be syllable-strings and # be interruption point[2]. These four types of repairs are represented as follows[3]:

- Repetition Repair: *A*#**A**
- Addition Repair: *A*#**BA**, *AB*#**ACB**
- Replacement Repair: *AB*#**CB**, *ABC*#**ADC**, *ABC*#**AC**, *AB*#**AC**
- Abandon Repair: *A*#**B**

The spoken corpus used in this paper consists of two commonplace, everyday conversations among friends. Each is about forty-minute long. There are four and five speakers in these two conversations, respectively. In total, this corpus contains 5,395 utterances and 2,602 turns. There are totally 448 self-repairs[4]. On the average, 17% of turns contain at least one repair. In this corpus, the repetition repairs have 71.65% of the repairs. Addition and replacement repairs have 11.16% and 9.82%, respectively. The rest (7.37%) are the abandon repairs. Because the repetition repairs and addition repairs form the majority (82.81%), we focus on them in this paper. Besides, since this paper corrects repairs based on acoustic and prosodic cues, the Chinese characters in the spoken corpus are converted into the corresponding syllables manually.

[2] The end of the repaired segment is called the interruption point. It is often accompanied by a disruption in the intonation contour.

[3] The repaired segment and the repairing segment are in *italic* and **boldface**, respectively.

[4] The speech repairs discussed in this paper are all self-repairs. That is, only the repairs accomplished by the same speaker are considered. This is because this kind of repairs is the most common form of repairs. Nevertheless, the present study includes repairs placed across different turns.

3. ANALYSIS METHODS

This section describes several cues for repair processing.

Cue 1: For repetition repair (*A*#**A**), the length of A (denoted as L_A) is limited to a certain range [2]. Similarly, for addition repair (A_1#$\mathbf{B}_1\mathbf{A}_1$, A_2B_2#$\mathbf{A}_2\mathbf{C}\mathbf{B}_2$), the lengths of A_1 (L_{A1}), A_2 (L_{A2}), B_1 (L_{B1}), B_2 (L_{B2}), C (L_C), are considered. This is an interesting problem in cognition. In our test, these are defined as follows:

$$1 \le L_A \le 4 \; ; \; 1 \le L_{A1}, L_{B1}, L_{B2}, L_C \le 3 \; ; \; 1 \le L_{A2} \le 2$$

Cue 2: In human conversation, most of the repairs occur between two consecutive utterances of one speaker without interrupting by other speakers. That is, if many utterances issued by other speakers are inserted between two utterances of the same speaker, the repairs usually do not occur. The spoken corpus shows this point.

- Total 13.69% (4%) of repetition (addition) repairs occur within the same utterance.
- Total 71.66% (84%) of repetition (addition) repairs occur between two consecutive utterances without interrupting by other speakers.
- Only 0.32% (0%) of repetition (addition) repairs occur across more than 3 utterances issued by other speakers.

Thus, when more than 3 (1) utterances pronounced by other speakers interrupt the speech of a speaker, we do not check whether there is a repetition (addition) repair or not. Besides, to limit the search range and prevent many false alarms generated by our system, we postulate there are no addition repairs within the same utterance.

Cue 3: In spontaneous or conversational speech, we find that there is a significant unfilled pause (silence) between a repaired segment and a repairing segment for the repetition repairs and addition repairs[5], whereas actual or intended repeated syllables usually do not have any unfilled pauses between them.

Cue 4: Glottal stop has the similar functions to unfilled pause. That is, a glottal stop may occur between the repaired segment and the repairing segment for the repetition repairs and addition repairs, whereas actual repeated characters usually do not have such a marker between them.

Cue 5: If two consecutive utterances are equal, repetition repairs usually do not occur within and

[5] Because the filled pauses such as um, un and er do not occur frequently in the spoken corpus, the effects of filled pauses are not demonstrated in this paper.

between them when the length of the utterances is long enough. This is because the matched syllable-string usually denotes an emphasis when it is long enough. For addition repairs, they always do not occur within or between two consecutive equal utterances no matter whether the length of these two utterances is long or short.

Cue 6: In Chinese conversation, some words or phrases are frequently repeated, but they are not repetition repairs. Typical examples are interjections (e.g., 哦 (o2, oh)) and phrase-final particles (e.g., 啊 (a5, a)). These patterns called type I cue patterns are used to increase the precision rate. That is, a repair is proposed when a string of syllables repeats, satisfies the criteria of cue 1, cue 2, cue 3 and cue 4, and the first syllable of the string does not belong to type I cue patterns.

In contrast to type I cue patterns, another kind of patterns, type II cue patterns, are also considered to increase the recall rate. That is, some repeated syllable strings (satisfy the criteria of cue 1 and cue 2) that do not satisfy the criteria of cue 3 and cue 4, but they are usually repetition repairs. Typical examples are pronouns such as 我 (wo3, I) and 你 (ni3, you). Based on type II cue patterns, some additional repairs can be proposed when a string of syllables repeats, it does not satisfy the criteria of cue 3 and cue 4, but the first syllable of the string belongs to type II cue patterns.

Similarly, these two types of cue patterns are also used in determining the addition repair. For repetition repairs (*A*#**A**), these two types of patterns are extracted from A. For type I addition repairs (*A*#**BA**), these two types of patterns are extracted from A. For type II addition repairs (*AB*#**ACB**), these two types of patterns are extracted from B.

Cue 7: According to the surface forms described in Section 2, we know replacement and addition repairs are two sides of a coin in some cases. The following three types of replacement repairs are usually proposed as the type I addition repair (*A*#**BA**).

(1) *AB*#**CB**,

(2) *ABC*#**ADC**,

(3) *ABC*#**AC**

Because the surface forms of (2) and (3) are more complex than that of type I addition repair, it is easy to distinguish them. That is, we can firstly identify these two types of repairs before the type I addition repairs. For the first case, we can use cue 9 to partially distinguish it from addition repair. However, we can expect that some false alarms will occur because of the ambiguities between these two patterns (*A*#**BA** and *AB*#**CB**).

Cue 8: For addition repair, the terminal pitch direction in repaired segment is important. When the terminal pitch direction of an utterance is rise or fall[6], the utterance is often finished by the speaker. In contrast, addition means an utterance is not finished. That is, the terminal pitch direction is often level. We use this cue to tell if there is an addition repair.

Cue 9: The word information is also considered. At the first sight, it seems to be contradictory. That is, before characters are identified and segmented further, how to know a string form a word. In our model, we generate a phonetic dictionary from lexical words. When a type I addition repair (X*A*#**BA**Y[7]) is proposed, we check if XA and AY can be found in this phonetic dictionary. If they are, they may be words. Thus this pattern is not regarded as a type I addition repair. For a proposed type II addition repair, i.e., X*AB*#**ACB**Y, we adopt the similar treatments.

4. EXPERIMENTAL RESULTS

The experiments is proceeded as follows. The test data is pipelined to the repetition module and then the addition module. For evaluating the performance, some conditions are listed below.

Condition 0 (C0): no cues

Condition 1 (C1): cue 1 + cue 2 + cue 3

Condition 2 (C2): cue 1 + cue 2 + cue 3 + cue 4

Condition 3 (C3): cue 1 + cue 2 + cue 3 + cue 4 + cue 5 + cue 6

Condition 4 (C4): cue 1 + cue 2 + cue 3 + cue 5 + cue 7 + cue 8

Condition 5 (C5): cue 1 + cue 2 + cue 3 + cue 5 + cue 6 + cue 7 + cue 8

Condition 6 (C6): cue 1 + cue 2 + cue 3 + cue 5 + cue 6 + cue 7 + cue 8 + cue 9

Based on these conditions, Tables 1 and 2 list the experimental results for correcting the repetition repairs and addition repairs. In these two tables, three metrics, i.e., precision, recall and P&R, are used for evaluating the system performance. The precision is the number of proposed repairs that were properly corrected compared to the number of total number of proposed repairs. The recall is the number of proposed repairs that were properly corrected compared to the total number of repairs. In addition to precision and recall scores, we also report Van Rijsbergen's F-Measure [7] which

[6] We classify the terminal pitch direction of an utterance into three types, i.e., rise (/), level (-) and fall (\).

[7] X and Y denote one or two syllables which immediately precedes and follows A, respectively.

Table 1 Results in Correcting the Repetition Repairs

	C0	C1	C2	C3
Precision	47.94%	84.14%	84.71%	93.87%
Recall	97.82%	76.01%	82.87%	90.65%
P&R	0.64	0.8	0.84	0.92

Table 2 Results in Correcting the Addition Repairs

	C0	C1	C2	C4	C5	C6
Precision	1.85%	26.72%	26.52%	36.17%	57.81%	76.09%
Recall	94%	70%	70%	68%	74%	70%
P&R	0.04	0.39	0.38	0.47	0.65	0.73

combines these two scores into a single measure. The F-measure (also called P&R) allows the differential weighting of precision and recall. With precision and recall weighted equally, it is computed by the following formula.

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

C0 in Tables 1 and 2 shows that the simple pattern matcher is not workable. The constraints C1 and C2 increase the precision, but some repairs cannot be captured. C3 has the best performance. In comparison to the "no cues" model, the precision rate for repetition repairs (addition repairs) increases from 47.94% (1.85%) to 93.87% (76.09%), while the recall rate decreases from 97.82% (94%) to 90.65% (70%).

From Tables 1 and 2, we find that glottal stop is a useful cue for repetition repair (compare C1 and C2), but it is not a useful cue for addition repair. Thus, this cue is not considered in C4, C5 and C6.

Although the cues for addition repairs increase the precision rate from 1.85% to 76.09%, it is clear that more cues are needed for correcting addition repairs.

5. CONCLUSIONS

Any spoken language systems will not perform well without treating speech repairs. Correcting speech repairs make more reliable environments for the subsequent processing. This paper employs acoustic and prosodic cues to correct the repetition repairs and addition repairs. Our algorithm assumes that the speech recognizer produces a sequence of syllables and the prosodic extractor produces some prosodic information.

O'Shaughnessy [8] claims that most speech repairs do not have lengthening prior to the hesitation pause. If this cue is used to correct repetition repairs, it can slightly increase the precision rate (95.37%), but the recall rate (76.95%) is greatly decreased.

Although our method can perform well in repetition repairs and addition repairs, replacement repairs and abandon repairs are not addressed in this paper. They have more complex surface forms and should be investigated further.

ACKNOWLEDGMENTS

We are grateful to Professor Kawai Chui for her kindly providing the spoken corpus to us.

6. REFERENCES

[1] P. Heeman and J. Allen "Detecting and Correcting Speech Repairs," *Proceedings. of ACL*, pp. 295-302, 1994.

[2] Y.S. Lee and H.H. Chen "Correcting Chinese Repetition Repairs in Spontaneous Speech," *Proceedings of ROCLING*, pp. 137-158, 1996.

[3] C. Nakatani and J. Hirschberg "A Speech-First Model for Repair Detection and Correction," *Proceedings of EUROSPEECH*, pp. 1173-1176, 1993.

[4] J. Bear, J. Dowding and E. Shriberg "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog," *Proceedings. of ACL*, pp. 56-63, 1992.

[5] K. Chui "Repair in Chinese Conversation," *Proceedings of International Symposium on Language in Taiwan*, pp. 75-96, 1996.

[6] Y.S. Lee and H.H. Chen "Applying Repair Processing in Chinese Homophone Disambiguation," *Proceedings of Applied Natural Language Processing*, pp. 57-63, 1997.

[7] C.J. van Rijsbergen *Information Retrieval*, Butterworths, London & Boston, 1975.

[8] D. O'Shaughnessy "Recognition of hesitation in Spontaneous Speech," *Proceedings. of ICASSP*, pp. 521-524, 1992.