

Acoustically verifying speech repair annotations

Fan Yang, Peter A. Heeman & Susan E. Strayer

Computer Science and Engineering
OGI School of Science and Engineering
Oregon Health & Science University, USA

Abstract

Identifying speech repairs is a critical part of annotating spontaneous speech. DialogueView is an annotation tool that provides visual and audio supports for directly annotating speech repairs. In this paper, we report the usability of *clean play*, a special feature implemented in DialogueView, which cuts out the annotated reparanda and editing terms and plays the remaining speech. We find that although *clean play* does not help users detect repairs, it does help them determine the extent of repairs. We also find that *clean play* improves users' confidence because they have another way to verify their annotations.

1. Introduction

The on-line nature of dialogues makes speech repairs prevalent. Speech repairs are a type of disfluency where speakers go back and modify or repeat something that they just said. Below is an example of a speech repair.

first *I need to find out* *um* *I need to get an engine*
 reparandum ↓ alteration
 editing term

Speech repairs tend to have a standard structure [8, 13], consisting of a reparandum, an optional editing term, and an alteration. The reparandum is the stretch of speech being replaced or cancelled. It is followed by the interruption point. The optional editing term consists of filled pauses (e.g. 'um') and cue words (e.g. 'I mean'). The alteration is the replacement for the reparandum. By removing the reparanda and editing terms, we arrive at the intended utterance of the speaker.

Speech repairs are a common phenomenon in spontaneous speech. Heeman & Allen [8] reported that 23% of speaker turns contain at least one speech repair, and 10% of the words in the Trains corpus [7] are in the reparandum or editing term of a speech repair. Shriberg [17] reported a higher disfluency rate of 57% in the Switchboard corpus.

Identifying speech repairs is a critical part of annotating spontaneous speech, as repairs impact utterance boundary and dialogue act coding decisions. Moreover, since utterances containing speech repairs are usually syntactically or grammatically ill-formed, repair annotations are also useful for training the language model of a speech recognizer to improve recognition rate [6, 18] and for building a parser for spontaneous speech [4, 5].

Although a number of tools can be used to directly or indirectly annotate speech repairs, DialogueView [9] provides better visual and audio supports for this task. DialogueView has a graphical means for annotating and displaying repairs, even embedded repairs. We are also experimenting with audio support, which is the subject of this paper.

By removing the reparandum and editing term, the intended utterance becomes syntactically well-formed at the interruption point (cf. [4, 10]). This "well-formedness" has

been used by several researchers. Bear et al. [2] used a two-step process in which pattern matching techniques first identify the reparanda and editing terms of potential repairs. The second step tests potential repairs by removing their reparandum and editing term, and seeing if the result is parsable. Kikui & Morimoto [12], as one source of evidence, judged whether the speech that precedes the reparandum can be syntactically followed by the alteration. The syntactic well-formedness was based on the part-of-speech tags. This technique was expanded upon by Heeman & Allen [8].

Just as the intended utterance is syntactically well-formed, it might also be intonationally well-formed. The prosody of the speech of the alteration might follow the prosody of the speech before the reparandum, just as if the reparandum and editing term had not been said. Hence, after users mark up a potential speech repair, they could listen to the intended utterance to help them decide the plausibility of the repair. We have built this *clean play* mechanism into DialogueView. We have personally found it helpful in choosing between alternative speech repair interpretations. To ascertain the usability of this *clean play* feature, we ran a controlled experiment. We find that although the *clean play* does not help in detecting repairs, it helps users in identifying the extent of repairs once they are detected. Also *clean play* improves users' confidence because they have another way to verify their annotations.

In section 2, we describe how speech repairs are annotated in DialogueView. In section 3, we describe the human-subject experiment in which we evaluate the *clean play*. In section 4, we give the conclusion.

2. Annotating repairs in DialogueView

Even though repairs are a normal part of spontaneous speech, annotation tools have yet to address them adequately. For example, Transcriber [1] allows word and utterance transcription, but has no direct means for annotating speech repairs. Mate workbench [15] can be used for annotating speech repairs only at the word-level.¹ It does not show the structure of speech repairs, especially the embedded ones.

DialogueView is a multi-level annotation tool. It can be used for annotating speech repairs, utterance boundaries, communicative status (such as overlapping, abandoned, incomplete, and uninterpretable), dialogue acts, and discourse structure. The interface of DialogueView consists of three views. The word view shows the exact timing of speech. The utterance view shows the dialogue as a sequence of utterances, as if it were a script for a movie. The intention view shows the dialogue as a hierarchy of discourse segment summaries and purposes. Two levels of abstraction are presented.

¹ The SRI annotation scheme [3] can be used for word-level repair annotation.

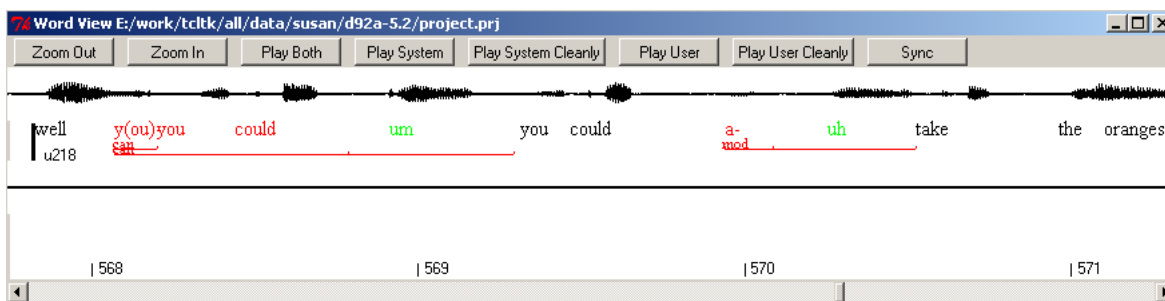


Figure 1: Interface of word view.

The utterance view abstracts away from the exact timing of the words and can even omit words that do not contribute to the content of the dialogue. The intention view abstracts away from the exact utterances that were said. Users get a general idea of what is happening in the dialogue from the higher level yet can view the lower level for details.

Users annotate speech repairs, along with utterance boundaries and communicative status, in the word view. Speech repairs should be annotated at the same time as utterance boundaries and communicative status since repairs have strong interactions with utterance segmentation and abandoned speech [9]. The word view takes as input the words said by each speaker and their start and stop times, and shows them time-aligned with the original audio signal. To annotate a repair, the user highlights a sequence of words and then tags it as a reparandum or as an editing term of a repair. The user can also specify the type of repair. Figure 1 shows how speech repairs are displayed in the word view. The words in the reparandum and editing term are underlined and displayed in a special color. Repairs can also be embedded. Figure 1 also shows an example where the speaker made a fresh start embedded in another fresh start.

In addition to visually displaying the scope of each repair, DialogueView also provides several audio playback options to help annotate speech repairs. Users can play each speaker channel individually or both combined (the Play Both, Play System and Play User buttons in Figure 1), which we refer to as *full play* since it plays everything that happened in the dialogue. Moreover, a special feature, *clean play* (the Play System Cleanly and Play User Cleanly buttons in Figure 1), is offered to let users hear the effect of their repair annotations. The *clean play* cuts out the stretch of speech annotated as reparandum and editing term and pastes the remaining speech together. If the repairs are correctly annotated, the *clean play* should sound fairly natural.

3. Evaluation of clean play

We conducted a human-subject experiment to investigate the usability of *clean play*. It was expected that people with the *clean play* feature would do better than people without it in annotating speech repairs.

3.1. Dialogue excerpts

Eight dialogue excerpts were taken from the Trains corpus [7]. Two were used for practice. The other six were used as material for the subjects to annotate speech repairs. Our experts annotated speech repairs for all eight dialogue excerpts. The last one proved too difficult, as there was a lot of self-talk. Hence it was excluded in the analysis of result. Table 1 shows details for these excerpts.

Table 1: Details of eight excerpts for coding speech repairs.

ID	Use	Number of repairs	Length
Tr1	demonstration	5	14 sec
Tr2	exercise	3	12 sec
1	subject coding	3	10 sec
2	subject coding	3	10 sec
3	subject coding	2	7 sec
4	subject coding	1	8 sec
5	subject coding	5	12 sec
6	subject coding	6	19 sec

3.2. Subjects

Thirteen subjects participated in the experiment. All were native English speakers. They were randomly divided into two groups: the control group had 3 females and 2 males, and the *clean* group had 4 females and 4 males. Subjects in the control group had only access to *full play*, which plays the original audio. Subjects in the *clean* group had the functionalities of both *full play* and *clean play*. We had more subjects in the *clean* group because we were interested in observing how people used the *clean play* function.

3.3. Experiment tool

A special version of DialogueView was built for subjects to annotate speech repairs. This special tool is self-contained, with instructions and exercises. Subjects were first taught the concept of speech repairs and how to code them using the annotation tool. Several examples were presented to familiarize subjects with speech repairs. The *clean* group subjects had the opportunity to listen to the intended utterances of the examples. Both groups of subjects were then given a dialogue excerpt (Tr1) and a list of steps to annotate the speech repairs. This gave subjects real experience in interacting with our tool, such as adding repairs, deleting repairs, and listening to the full and intended utterances (*clean* group only). The final phase of training is an exercise (Tr2) in which subjects annotated speech repairs on their own and then compared their annotation with our expert annotation.

After the training, subjects were given the six dialogue excerpts to annotate one by one in the same order. The tool prohibits subjects from going back to previous excerpts. Subjects' interactions with the tool, including adding and deleting speech repairs, and audio playback with *full play* and *clean play*, were all automatically logged.

3.4. Procedure

Subjects completed the training and annotation by themselves in a private room without any interference. They could call the tester at any time to answer questions about using the tool.

After the experiment, subjects filled in a questionnaire to give their feedback, such as degree of confidence on their annotation.

3.5. Results

Our expert annotation serves as the gold standard for evaluating subjects performance. We adopt *detection* and *correction* to evaluate subjects performance.¹ We use a less restricted definition of detection than what is typically used. If the reparandum or editing term of a repair annotated by a subject overlaps the reparandum or editing term of a repair in the gold standard, we say that the gold standard repair is detected by the subject. A repair in the gold standard is missed if it is not detected by the subject. If the words in the reparandum and editing term of a gold standard repair are the same as the words in the reparandum and editing term that the subject annotated, we say that this repair is corrected by the subject. A corrected repair implies that it is detected. A detection means the subject was aware of a disfluency and a correction means the subject located the extent of the disfluency. Figure 2 shows some examples of detection and correction.

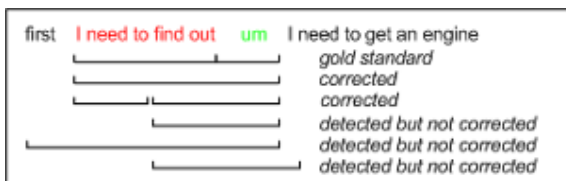


Figure 2: Examples of detection and correction.

Figure 3A shows the performance of both groups in detecting each of the 14 repairs. Overall, the mean detection rate for the control group is 88.6% (62/70), and for the clean group is 89.3% (100/112). Statistically, we don't see a significant difference in the detection rate between the two groups (pairwise signtest, $p = 1$). This is not surprising because we expect *clean play* to help correct speech repairs, not detect them, as we will explain in the discussion.

Figure 3B shows the performance of both groups in correcting each of the repairs. Overall, the mean correction rate for the control group is 67.1% (47/70), and for the clean group is 72.3% (81/112). This suggests that the clean group subjects are doing a little better than the control group in correcting repairs.

Due to the small performance improvement, we also examine how *clean play* was used. First, we find that all clean group subjects verified every repair they annotated with *clean play*. Second, in nine cases, they changed their annotation after using *clean play*. In eight of the cases, they changed from a wrong corrected repair (but correctly detected) into a correction. In the ninth case, a subject deleted a correct annotation. Hence, the clean group improved their rate in correcting repairs from 66.0% (74/112) before using *clean play* to 72.3% (81/112). This suggests that it is the use of *clean play* that accounts for the improvement of the clean group over the control group. Overall, *clean play* reduced the correction errors of detected repairs from 26.7% (27/101) to 19.0% (19/100), giving a relative improvement of 28.8% in correcting repairs.

¹ We report recall only. From our data we only see three cases of false positive, two in the clean group and one in the control group. We believe that those are minor and can be ignored.

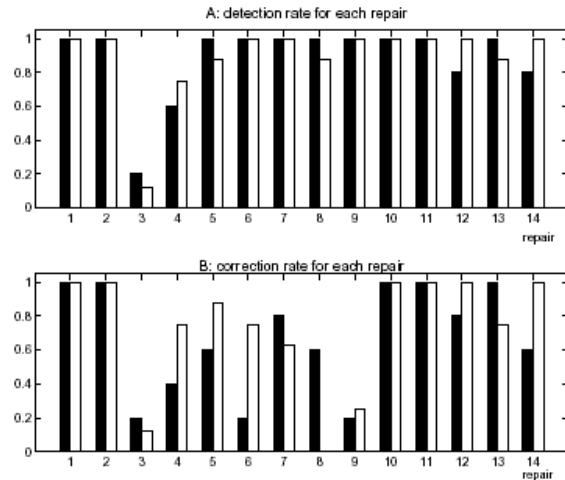


Figure 3: Performance on each repair: black for the control group and white for the clean group.

There are a couple of repairs where the control group did better than the clean group. For repair 3, although the control group has a higher correction rate, the correction rate over detected repairs is almost the same for both groups. This means that the difference in repair correction rate is due to the difference of subjects' capability in detecting repairs, instead of the use of *clean play*. The utterance of for repair 8 is "E two and E three *are boths* want to leave Elmira at the same time", where the reparandum and editing term are in italics followed by the repair number as a subscript. Both groups have similar detection rate, but vastly different correction rates. In fact, the clean group mistakenly thought that the interruption point was after the word "are" rather than the word "both" which is followed by a short silence. More work is needed to explain this negative result.

Repairs 7 and 13 show limitations of *clean play*. The utterance of repair 7 is "yes *the7* the problem here is that ..." If subjects mistakenly coded the words "yes the" as the reparandum, the remaining speech still sounds fluent. The *clean play* does not help in finding this mistake. The utterance for repair 13 is "how long would it take *to get13* to take *engine well let's see14* engine number two..." One clean group subject annotated it as "how long would it take *to get13* engine *well let's see14* engine number two..." This annotation under *clean play* sounds as good as the correct annotation. Just as a spell checker can not distinguish between "out" misspelled as "our", *clean play* can not catch all incorrect annotations.

We also asked subjects for their degree of confidence on their annotation (on a scale from 1 to 5). The clean group subjects reported higher confidence ($mean = 3.5$; $\sigma = 0.53$) than the control group ($mean = 2.8$; $\sigma = 0.84$). Our clean group subjects were satisfied with the *clean play*: they ranked the *clean play* as useful as the *full play*. Interestingly, one subject in the control group mentioned that he would like to listen to the intended utterance.

3.6. Discussion

It is not surprising that our tool does not help in detecting speech repairs because there are strong acoustic cues around the interruption point. Levelt & Cutler [14] reported the correlation between error repairs (repairs of erroneous information) and increased intonational prominence at the beginning of an alteration. This result was confirmed by Howell & Young [11]. They found that some repairs tend to

have a pause around the interruption point and have a strong accent at the onset of alteration. Nakatani & Hirschberg [16] found that the reparanda often end in word fragments (73.3%) and are often accompanied with glottalization and coarticulation, especially for those ending in fragments. They also found that filled pauses and cue phrases occur significantly more often in non-fragment repairs than in fragment repairs. These cues can be heard in *full play*, which subjects in both groups had access to.

The *clean play* makes the assumption that the intended utterance should sound “fluent”. Although a strong accent at the onset of alteration is found at some repairs, many repairs do not have this feature [11, 14]. Our positive results suggest a thorough investigation of prosodic cues between the speech before the reparandum and the onset of the alternation is warranted.

Our results show that overall people with the *clean play* do a little better than people without it in correcting speech repairs. When users have access to a transcription of the words, including word fragments, giving them *clean play* only gains a modest improvement. This is because a lot of repairs can be detected and corrected by just looking at the words. To get results that are statistically significant, a much larger sample size is needed. An area that we have not investigated is the advantage of our tool when multiple repairs occur in a short stretch of speech. The *clean play* will play the effect of current annotation, hopefully allowing the user to catch the remaining repairs.

4. Conclusion

In this paper, we described our annotation tool, DialogueView, which provides visual and audio support for annotating speech repairs. We find that although our *clean play* feature, which plays the speech left after cutting out speech repair reparandum and editing term, does not help people detect repairs, it does help people identify the extent of repairs, reducing their error rate by 28.8%, and improves their confidence in their speech repair annotations.

5. Acknowledgements

The authors acknowledge funding from the Intel Research Council. We also thank members of CSLU and CHCC for helpful discussions.

6. References

- [1] Barras, C., E. Geoffrois, Z. Wu, & M. Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, vol. 33(1–2), pp. 5–22.
- [2] Bear, John, John Dowding & Elizabeth Shriberg. 1992. Integrating multiple knowledge sources for the detection and correction of repairs in human–computer dialogue. *Proceedings of 30th ACL*.
- [3] Bear, John, John Dowding, Elizabeth Shriberg & Patti Price. 1993. A system for labeling self-repairs in speech. *Technical Report 522*, SRI, February 1993.
- [4] Charniak, Eugene & Mark Johnson. 2001. Edit detection and parsing for transcribed speech. *Proceedings of 2nd NAACL*, 2001.
- [5] Core, Mark G. & Lenhart K. Schubert. 1999. A syntactic framework for speech repairs and other disruptions. *Proceedings of 37th ACL*.
- [6] Heeman, Peter A. 1999. Modeling speech repairs and intonational phrasing to improve speech recognition. In *Automatic Speech Recognition and Understanding Workshop*, Keystone Colorado, December 1999.
- [7] Heeman, Peter A. & James F. Allen. *The Trains spoken dialogue corpus*. CD-ROM, Linguistics Data Consortium, 1995.
- [8] Heeman, Peter A. & James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers’ utterances in spoken dialogue. *Computational Linguistics*, vol. 25(4).
- [9] Heeman, Peter A., Fan Yang & Susan E. Strayer. 2002. DialogueView: A dialogue annotation tool. *Proceedings of 3rd SIGDial workshop on Dialogue and Discourse*, Philadelphia.
- [10] Hindle, Donald. 1983. Deterministic parsing of syntactic non-fluencies. *Proceedings of 21st ACL*.
- [11] Howell, P. & K. Young. 1991. The use of prosody in highlighting alteration in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology*, vol. 43A(3).
- [12] Kikui, Gen-ichiro & Tsuyoshi Morimoto. 1994. Similarity based identification of repairs in Japanese spoken language. *Proceedings of 3rd ICSLP*.
- [13] Levelt, Willem. 1983. Monitoring and self-repair in speech *Cognition*, vol. 14, pp. 41–104.
- [14] Levelt, Willem & Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, vol. 2, pp. 205–217.
- [15] Mckelvie, D., A. Isard, A. Mengel, M. B. Moeller, M. Grosse & M. Klein. 2001. The MATE Workbench – An annotation tool for XML coded speech corpora. *Speech Communication*, Special issue, “Speech Annotation and Corpus Tools”, vol. 33(1–2), pp. 97–112.
- [16] Nakatani, Christine H. & Julia Hirschberg. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, vol. 954, pp. 1603–1616.
- [17] Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proceedings of 4th ICSLP*.
- [18] Stolcke, Andreas, Elizabeth Shriberg, Dilek Hakkani-Tür & Gökhan Tür. 1999. Modeling the prosody of hidden events for improved word recognition. *Proceedings of 6th Eurospeech*, 1999.