

On not Recognizing Disfluencies in Dialogue

R. J. Lickley and E. G. Bard

Human Communication Research Centre and Department of Linguistics
University of Edinburgh

ABSTRACT

This paper tests the hypothesis that listeners miss disfluencies or fail to transcribe them accurately because disfluencies interfere with the normal relationship between speech sound and linguistic context in human spoken word recognition. In a word-level gating experiment 16 listeners heard a total of 56 disfluent utterances selected from a corpus of spontaneous speech, 56 length-matched fluent controls, and 56 fluent foils. The proportion of words never recognized was greater in disfluent utterances than in controls. The failures clustered around the point where the disfluency interrupted the utterance, occurring particularly within the reparanda, but were not found at corresponding locations in uninterrupted controls. Repetition disfluencies, where pre- and post-interruption portions might easily be construed together, allowed more successful word recognitions than recast disfluencies, where reconstruction of a single intended utterance would be difficult, if not impossible. The results have implications both for understanding human speech recognition and for improving the robustness of ASR systems.

1. INTRODUCTION

Anyone who has ever attempted a verbatim transcription of spontaneous speech knows that disfluent portions are much more difficult to transcribe accurately than fluent sequences. Disfluencies are often missed in the transcription process and, when spotted, prove difficult to resolve into words. For speech presented at normal speed, instructions to attend carefully to disfluencies increase bias to report them but not accuracy in locating them [11]. On the other hand, listeners can locate the onset of a disfluency promptly when speech is interrupted to solicit their judgments [8, 9], and the contents of the disfluency affect the processing of subsequent fluent speech [6]. It seems that disfluent speech is processed, but remains evanescent.

Inconvenient though this phenomenon is for transcribers, it is one which ASR systems might wish to emulate. ASR systems have considerable difficulty in discarding reparanda, that is, those parts of disfluencies which need to be expunged before a fluent sequence can be reconstructed. Much attention has

been devoted to specifying the distinctive acoustic or structural features of disfluencies which would reliably trigger an accurate editing process [3, 5, 12, 14]. We are currently investigating the human perceptual solution to this problem.

We propose that people do not mentally transcribe and then expunge disfluent speech. Instead, they fail to recognize the acoustic material of disfluencies as words or they recognize it with so much delay that portions of the speech will be lost from memory as new input arrives. We make this prediction because the normal processes of word recognition in running speech ought to be severely disrupted by the interruptions which disfluencies create.

Normally, listeners depend on both preceding and subsequent context to recognize words in running speech [2, 4, 7]. While most words can be recognized as soon as they are heard with their prior contexts, some remain indecipherable until a prosodic or constituent boundary occurs up to several words later in the utterance [13]. The more prior context a word has, that is, the later in the sentence it occurs, the more likely immediate recognition is. When disfluencies interrupt speech, they disrupt both contexts on which listeners depend. By creating shorter sequences of words which can be construed together, disfluent interruptions reduce the extent of supporting prior context and create conditions where later material should be important to the recognition process. But by truncating reparanda before prosodic or constituent boundaries, they also remove or delay those sites where late recognition would normally occur.

We report an experiment which tested specific predictions arising from this view by comparing listeners' recognition of words in naturally occurring disfluent utterances and in length-matched fluent utterances. To maximize rates of recognition, materials were presented via word-level gating. This method increments the presented portion of an utterance by one word in each successive trial.

First, we predict more failures to identify words from disfluent items than from fluent items with the same total length. Second, the difficulties should cluster around the interruption point, the point where the reparandum ends and the

rest of the utterance begins. The greatest rate of loss should be in reparanda, where subsequent context is truncated. Following the interruption point, prior context is initially minimal and the disfluent resumption should not support timely recognition as well as the uninterrupted control. Finally, we predict an effect of the type of disfluency. If the disfluency recasts the false start of an utterance, then material following the interruption point may not be construable with preceding words prosodically or syntactically: the context which would permit late recognition may never arrive and the continuation itself will lack prior context. If the disfluency is a repetition, then the pertinent later context is merely postponed by a few words. Repetitions should therefore support more successful word recognitions.

2. METHOD

2.1. Materials and Design

All speech materials were spontaneous utterances from the HCRC Map Task Corpus [1], a collection of 128 digitally recorded dialogues between pairs of Scottish undergraduates engaged in a route communication task. Dialogues took place in a recording studio. Each speaker was recorded by close-talking microphone on a separate channel. Disfluencies were produced in the course of complex interactions.

Disfluent utterances from the corpus were classified prior to the experiment as repetitions, insertions, deletions, recasts, or as more than one of these, and coded for the number of words in the *reparandum* and in the *repair*, the words which ‘overwrite’ the reparandum. The *interruption point* separates reparandum and the repair. Classification and transcription were performed and word onsets and offsets marked with the aid of Entropic speech editing and spectrographic facilities.

Twenty-eight disfluent utterances were selected which included verbatim repetitions but no other kind of disfluency. Another 28 disfluent items were selected which matched the repetitions in reparandum length and recast earlier portions of the utterance. Half of each group ended in whole words, half in word fragments. Examples of each follow, with IP marking the interruption point.

1. REPETITION: Right, there’s a IP there’s a line about a quarter of the way down.
2. RECAST: Well, until y- IP stop at the ‘B’.

Each disfluent utterance was paired with a fluent utterance of the same length in words produced by the same speaker. These 112 test utterances and 56 fluent fillers were distributed among 4 tapes by Latin square and blocked by speaker ($N = 11$).

fluent utterances disfluent utterances

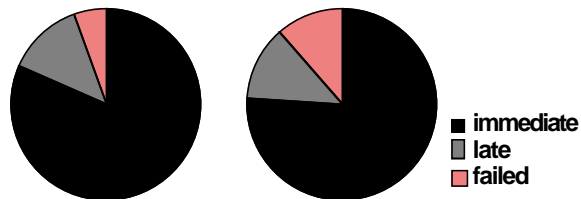


Figure 1: Distribution of outcomes of all attempts ($N = 4384$) at recognizing words in disfluent items and their fluent controls.

2.2. Subjects

Subjects were 16 members of the Edinburgh University community, all native speakers of English with no known hearing loss. Four subjects heard each tape. Each subject heard all 56 filler utterances and 56 test utterances, 14 each from each cell of the design.

2.3. Procedure

Subjects were told that they would hear utterances beginning with their first word and then including one additional word on each trial until the utterance was complete. Their task was to identify each new word as soon as they had heard it, writing it on an answer sheet which allowed one block for each word presented on each trial. They were encouraged to guess and allowed to change their transcription for any word on the line corresponding to the trial when they changed their mind, but not to alter previous lines.

3. RESULTS

Two faulty recast items and their fluent controls were discarded, leaving 14 repetitions and 12 recasts. Over all subjects and materials, the data comprise 4384 attempts, usually over multiple trials, to recognize spoken words, half in fluent and half in disfluent utterances. A word received an *immediate recognition* if correctly identified by a subject on its first presentation with only prior context, a *late recognition* if first recognized after at least one additional word, and a *failed recognition* if never correctly transcribed.

Figure 1 shows that, as predicted, words in disfluent utterances are the more difficult to recognize ($\chi^2_{(4384)} = 48.82$, $df = 2$, $p < .0001$). Disfluent items yielded fewer immediate recognitions than fluent (76.1% v 81.7%) and more failures (11.5% v. 5.6%), while late recognitions occurred at a similar rate in the two (12.4% v 12.7%). The largest component of χ^2 was contributed by the difference in rates of failure (2×22.2).

As predicted, also, difficulties clustered around the interruption point. To test this proposal, each disfluent utterance was divided into 4 parts. The *reparandum* immediately preceded the interruption point. Any words preceding the reparandum

were classed as the *original utterance*. Immediately following the *interruption point*, a *repair* was delimited in all disfluent utterances, including either a genuine replacement for the reparandum or, if none existed, a string of words equal in length to the reparandum. The remaining words comprised the *continuation*. For purposes of comparison, fluent utterances were divided at the same points as their respective disfluent counterparts.

Figures 2a - 2d display the distributions of recognition outcomes for disfluent and fluent items within each part of the utterance. All four comparisons show significant differences. The most marked are found in reparanda ($\chi^2_{(768)} = 84.00$, $df = 2$, $p < .0001$). As we would predict from their truncated subsequent contexts, disfluent reparanda (Fig. 2b) produce many more failures (26.8% v 3.6%) than fluent and fewer late recognitions (9.1% v 17.7%). The excess failures contribute the largest component of χ^2 (2×33.9). Disfluent repairs (Fig. 2c) are also difficult to recognize ($\chi^2_{(864)} = 46.16$, $df = 2$, $p < .0001$). With effectively truncated prior context, these show more failures (14.8% v 4.2%), fewer immediate recognitions (64.4% v 83.6%), and more late recognitions (20.8% v 12.3%) than the corresponding parts of fluent controls. Again the disproportionate rate of failures makes the major contribution to χ^2 . For beginnings and ends of the utterances (Fig. 2a, 2c), the effects of disfluency are less extreme. Disfluent original utterances, with abbreviated subsequent contexts, produce fewer late recognitions than their fluent counterparts (11.2% v 17.5%: $\chi^2_{(1144)} = 14.20$, $df = 2$, $p = .0008$). Finally, continuations yield more late recognitions than are needed in the final portions of their fluent controls (10.3% v 7.1%: $\chi^2_{(1608)} = 8.04$, $df = 2$, $p < .02$).

Figure 3 shows how closely the fluency differences are linked to the interruption point. It displays the proportions of immediate, late and failed recognitions for 4 words on either side of the interruption point. Immediate recognitions fall off just after the interruption point in disfluent utterances. Just before it, the rate of failures peaks abruptly. Just after it, the rate of late recognitions rises almost to the level found at the first word of the utterance. Fluent utterances produce no such effects at these points. Multiple regression equations bear out this picture. Separate equations were used to predict proportion of each outcome over all the words in reparanda and all following the interruption point, with word duration, distance from start of utterance, distance from end of utterance, and distance from interruption point as concurrent independent variables. For disfluent utterances only, and with all other position indices statistically controlled for, rate of immediate recognitions falls significantly in the reparandum as the interruption point approaches (Multiple $R^2 = 0.16$; $F = 18.40$; $df = 4, 379$; $\beta = -0.28$, $t = -5.36$, $p < .01$) and rises as it is left further behind (Multiple $R^2 = 0.06$; $F = 13.48$; $df = 4, 799$; $\beta = 0.11$, $t = 2.31$, $p < .02$). The failure rate rises significantly in disfluent utterances as the interruption point approaches (Multiple $R^2 = 0.14$; $F = 15.25$; $df = 4, 399$; $\beta = 0.30$, $t = 5.73$, $p < .01$).

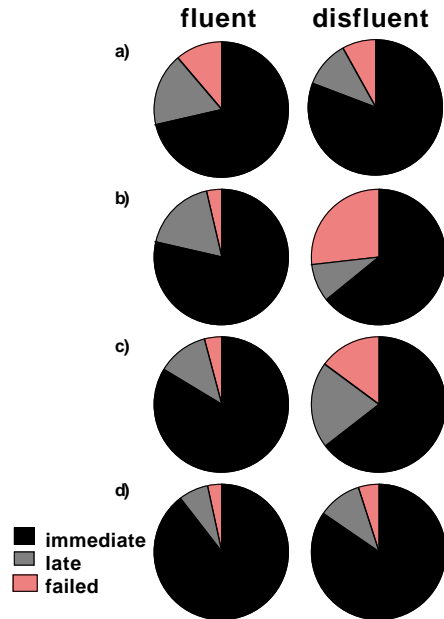


Figure 2: Distribution of outcomes of all attempts at recognizing words in disfluent items and their fluent controls by part of utterance: a. Original utterance; b. Reparandum; c. Repair; d. Continuation

Finally, recognition outcomes depend on the relationship between what precedes and what follows the interruption point. Repetition disfluencies, where the two are more likely to be parsable as a single sequence once extra tokens of repeated words are removed, are more successfully recognized than recast disfluencies, where reconstituting an utterance should be more difficult. As Figure 4 shows, recasts produce more failures to recognize words (13.3% v 9.9%; $\chi^2_{(2192)} = 6.86$, $df = 2$, $p < .04$).

4. DISCUSSION

Disfluencies certainly do disrupt the recognition process, even when we interrupt presentation of spontaneous speech frequently and present substrings of utterances many times. At the beginning of disfluent utterances, expected late recognitions are blocked. Reparanda, which need to be removed to reconstruct fluent utterances, are most susceptible to failures to identify words correctly. Repairs, allow less instantaneous recognition and more ultimate errors than the corresponding parts of fluent utterances. By the ends of disfluent utterances, well beyond the interruptions, recognition is still unusually delayed. In general, then, disfluent utterances yield erroneous identifications or delayed success.

The results make a prima facie case for the relevance of normal word recognition processes to human treatment of disfluencies. They do not yet provide a complete account of a mechanism for dispensing with what speakers intended to

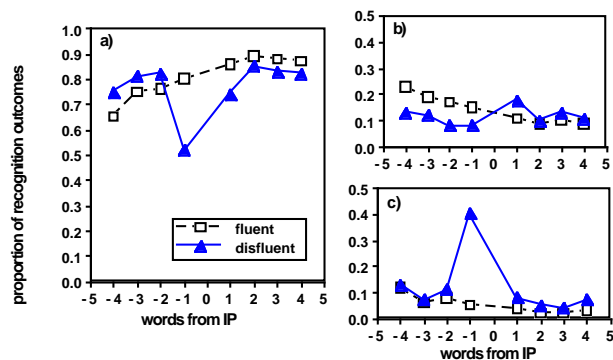


Figure 3: Comparison of fluent and disfluent recognition outcomes by distance from interruption point: a. Immediate recognitions; b. Late recognitions; c. Failed recognitions.

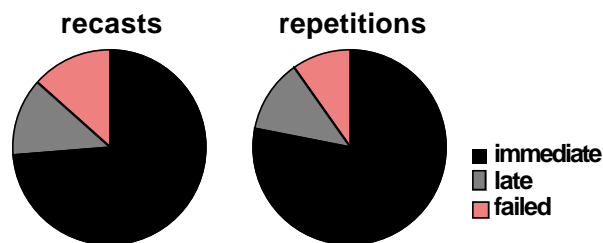


Figure 4: Distribution of outcomes of all attempts at recognizing words in recast ($N = 1004$) and repetition ($N = 1188$) disfluencies

discard. Failures to recognize a word correctly may be due to the absence of an internal account of the word, to the listener's inability to select an account or to the presence of the wrong account. It may be that these are effectively the same thing, the signs of a processing burden imposed by blocking the normal contribution of higher level information to the recognition of words in running speech. We know that word recognition involves selecting one out of a number of lexical hypotheses for the identity of a stretch of speech [10]. With insufficient contextual information, no clear winner may emerge from the set of competing hypotheses. When not forced to respond, listeners can simply delay their choice, maintaining strings of word hypotheses for later resolution [2]. During disfluent utterances, listeners would accumulate a considerable processing load in the form of unresolved analyses, untenable analyses, and reanalyses. At various times it may become impossible to retain all the information needed to continue with unresolved tasks. Parts of utterances may have to be abandoned without providing any word-level account of them. Whether listeners finally omit just those areas where the current experiment found recognition difficulties remains to be discovered by further experimentation. Whether the kind of architecture which creates these difficulties for people can be constructed for ASR systems remains to be seen.

5. REFERENCES

1. A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366, 1991.
2. E. G. Bard, R. Shillcock, and G. T. M. Altmann. The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context. *Perception and Psychophysics*, 44:395–408, 1988.
3. J. Bear, J. Dowding, and E.E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the ACL*, pages 56–63, 1992.
4. C. Connine, D. Blasko, and M. Hall. Effects of subsequent sentence context on auditory word recognition: temporal and linguistic constraints. *J. Memory and Language*, 30:234–250, 1991.
5. D. Duez. Acoustic correlates of subjective pauses. *J. Psycholinguistic Research*, 22:21–39, 1993.
6. J. Fox Tree. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *J. Memory and Language*, 34:709–738, 1995.
7. F. Grosjean. The recognition of words after their acoustic offset: evidence and implications. *Perception and Psychophysics*, 38:299–310, 1985.
8. R. Lickley and E. Bard. Processing disfluent speech: recognising disfluency before lexical access. In *Proceedings of ICSLP*, pages 935–8, 1992.
9. R. Lickley, E. Bard, and R. Shillcock. Understanding disfluent speech: is there an editing signal? In *Proceedings of ICPHS*, volume 4, pages 98–101, 1991.
10. W. Marslen-Wilson. Activation, competition, and frequency in lexical access. In G. Altmann, editor, *Cognitive models of speech processing*, pages 148–172. MIT Press, Cambridge, MA, 1990.
11. J. Martin and W. Strange. The perception of hesitation in spontaneous speech. *Perception and Psychophysics*, 3:427–38, 1968.
12. C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95:1603–1616, 1994.
13. R. Shillcock, E. G. Bard, and F. Spensley. Some prosodic effects on human word recognition in continuous speech. In *Proceedings of SPEECH '88*, pages 819–26, 1988.
14. E.E. Shriberg and R.J. Lickley. Intonation of clause-internal filled pauses. *Phonetica*, 50:172–179, 1993.