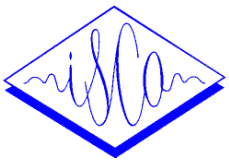

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PROCESSING DISFLUENT SPEECH: RECOGNISING DISFLUENCY BEFORE LEXICAL ACCESS

R.J. Lickley †* and E.G. Bard *‡

† Centre for Speech Technology Research, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK
email: robin@cstr.ed.ac.uk
* Department of Linguistics, University of Edinburgh
‡ Human Communication Research Centre, University of Edinburgh

ABSTRACT

As work on speech understanding moves towards the study of spontaneous rather than carefully prepared read speech, the problems posed by disfluency need to be addressed. The first problem for the processor is to detect that a disfluency has occurred. Previous experiments [11] have shown that listeners are usually able to detect disfluency within one word of the interruption. This paper presents results of a further experiment which looks more closely at recognition points of disfluency and of the following word. It is found that listeners are able to detect that disfluency has occurred soon after the onset of the following word and prior to recognition of the word itself. Taken together with the results of an experiment with low-pass filtered speech [12], the results suggest that prosodic information may play a key rôle in the processing of disfluent speech.

10.21437/ICSLP.1992-286

1 INTRODUCTION

With the vast majority of studies on language processing and speech perception being based on written language or carefully prepared read speech, the question of fluency and particularly how to handle hesitation and disfluency has not arisen until fairly recently. The experiment described in this paper is the latest in a series of perception tests which look at aspects of human processing of disfluent sentences taken from spontaneous English conversations.

For the purposes of this discussion, we take the terms "disfluency" and "repair" to be interchangeable and to refer to repetitions and false starts of lengths varying from less than a syllable to several words.

Example 1: Repetition:
'And you'd re- Ξ you'd really need about eight ...'

Example 2: False Start:
'Because although the bell Ξ the rules say that ...'

We refer to the part of the utterance following the *interruption* (Ξ) as the *continuation* following Levelt ([9]).

Disfluency occurs with great frequency in spontaneous speech. Various authors with different corpora describe the frequency of occurrence in different ways: Levelt [9] finds one repair for every three descriptions of simple patterns; Blackmer and Mitton [3] found repairs every 4.8 seconds; the corpus used for the present study contains some form of disfluency (including "ungrammatical" silent pauses) on average approximately every seven words. The result of this is that, in attempting to understand spontaneous speech, the processor is very frequently faced with input containing apparently ungrammatical text. Clearly, then, disfluency presents a major processing problem for both psychological and computational models of speech perception.

Despite its prevalence, everyday experience tells us that human listeners are often unaware of the occurrence of disfluency. This suggests that the human speech processing mechanism has early access to any cues that are available in the speech signal. This paper addresses the questions of what cues are available and how soon listeners are able to use them.

Very little work in psycholinguistics has so far addressed these problems. In the search for cues, Howell and Young [8] suggest that pauses and added stress on the first word of the continuation are used by listeners in processing disfluent speech. Their experiments, using synthesised speech with artificial repairs, suggest that a 200msec pause and added stress, in the form of "a loudness increase and durational change corresponding to a primary stress" on the first word of the continuation, are helpful to listeners in processing repairs involving alterations (but not repetitions). However, the importance of pause length and added stress (or markedness) as cues to processing disfluent speech can be put into perspective by examining data on their frequency of occurrence from speech production research. Blackmer and Mitton [3] find that 48.6% of overt repairs in their corpus have cut-off-to-repair times of less than 100ms and 19.2% have times of 0msec. Furthermore, spontaneous speech contains frequent instances of mid-clause silent pauses. So the pause can not be said to be a very reliable cue to repair. Cutler [5] and Levelt and Cutler [10] find that *prosodic marking* in repairs occurs most commonly in lexical repairs (38% of lexical repairs being marked in [5] and 45% in [10]) and particularly in error as opposed to appropriateness repairs. They conclude that such marking is used by the speaker for contrastive accentuation and not as a specific marker of disfluency.

However, the experiment using low-pass filtering on spontaneous false starts and repetitions reported in [12] does suggest that some prosodic factors (other than contrastive stress) are important in helping listeners recognise speech as disfluent.

Within computational linguistics, the problem of processing disfluencies is approached mainly from a syntactic angle. Hindle's algorithm [7] relies on the detection of a discrete phonetically identifiable *editing signal* and then uses a series of syntax-based editors to extract a parsable sentence. Bear, Dowding and

Shriberg [2] point out that such an editing signal has yet to be found and therefore take a different approach to identifying potential locations of disfluencies. They use a word- and syntax-based pattern matching technique to identify possible disfluencies before applying information from subsequent syntactic, semantic and acoustic analyses to distinguish true disfluencies from false positives. F0 values and pauses are found to be of use in the acoustic analyses.

The question of *when* during the processing of an utterance the processor has enough information to detect disfluency is not really relevant for computational models, which do not perform on-line processing and assume the availability of syntactic information on both sides of the interruption. In psycholinguistics, Lickley, Bard and Shillcock [11] have found that listeners are usually able to recognise disfluency within one word of the disfluent interruption but not before the onset of that word (ie subjects did not detect an editing signal prior to the onset of the continuation).

The results of the word-level gating experiments described in [11] left open the question of what information subjects used in detecting disfluency. Since the word following the interruption was recognised at first presentation in around 80% of cases (not an unusually low or high rate, [1]), it is possible that subjects made use of syntactic knowledge in their fluency judgements.

The experiment described in this paper uses 35msec gating to find more precise recognition points for disfluencies in the same materials. The experiment also allows us to determine whether listeners are able to recognise the first word of the continuation and therefore have access to lexical and syntactic information before they can detect disfluency or if they are able to detect disfluency before lexical access. It is found that in most cases listeners are able to detect disfluency before they have recognised the word immediately following the interruption and that they can therefore use information other than syntactic in detecting disfluency.

2 A 35MSEC GATING EXPERIMENT

2.1 Introduction

This experiment was designed to find recognition points for disfluencies within the first word following the interruption for a selection of disfluent utterances used in previous experiments. The previous experiments had established that subjects were usually able to detect disfluency by the offset of the crucial word but not prior to its onset [11]. A further purpose of this experiment was to find out when recognition of disfluency took place with respect to the recognition point of the crucial word. It is the latter question that we focus on in this paper.

2.2 Materials

The test materials were a set of utterances taken from a corpus of 6 studio-recorded spontaneous dialogues. Twenty-eight disfluent utterances (containing repetitions and false starts of various lengths) were chosen as representative of the distribution of the types of disfluency found in the whole corpus. Twenty-eight fluent control utterances were selected from the same corpus, matching the disfluent utterances for structure, length and prosody as far as possible. Rehearsed fluent versions of all the spontaneous utterances, produced by the same speakers, were also used as controls, making a total of 112 utterances for the whole experiment (the method used to produce the rehearsed utterances is described in [11]). All the speech material used in the experiment was sampled at 20kHz through an 8kHz filter.

The materials were prepared for presentation to 4 subject groups. The four sets of materials (spontaneous disfluent and fluent and their rehearsed versions) were blocked by speaker, organised by latin square and then randomised to decide the order of presentation. As a result, each subject group heard 5 utterances from 4 speakers and 4 from 2 speakers and heard a total of 7 members of each set of materials.

2.3 Procedure

The experiment was preceded by a taped introduction with full instructions and examples and a practice test. There was then a pause for the practice test to be checked and for subjects to ask questions.

Before the test items for a new speaker were presented, a short passage of conversation involving that speaker was heard, to help subjects get accustomed to the voice. Each test item consisted of three phases: about ten seconds of the prior conversation, for discourse orientation; the beginning of the test utterance, up to the moment prior to the crucial words; the gated presentation, which included the beginning of the test utterance (ungated) on each presentation. The words gated were the word prior to and the word following the interruption point in the disfluent cases and the 2 words at the equivalent point in the control utterances.

Gating commenced at the onset of the word prior to and continued until the offset of the word following the interruption. Gating was in increments of 35msec. The first stimulus for an item consisted of the beginning of the utterance up to the moment prior to the first crucial word, the second stimulus contained the first stimulus plus 35msec of the word and so on, each stimulus increasing in length by 35msec until the offset of the second crucial word.

Sufficient time was allowed between each presentation for subjects to write their responses and tones preceded the onset of each stimulus.

The experiment was run in two sessions of about 45 minutes.

2.3.1 Tasks

There were two tasks to be completed at each gated presentation: word recognition and fluency judgement.

Word Recognition

Subjects were instructed to write down what they thought the current word was at each gated presentation and to make any amendments required to previous judgements in the appropriate part of the answer sheet, without erasing earlier erroneous judgements. They were asked to try to guess a whole word where possible, rather than giving gradual transcriptions.

Fluency Judgement

Subjects were asked to make a judgement on a scale of 1-5 as to the fluency of the utterance at the latest gated presentation. (1 signified "fluent", 5 "disfluent" and 3 "don't know"). The judgement was marked on the answer sheet alongside the word judgement, by circling one of the printed numbers 1-5.

2.4 Subjects

Subjects were 43 native speakers of English, members of the University community (three groups of 11 and one of 10). They were seated in sound-proof booths and listened to the digital

tapes through high-quality headphones.

2.5 Results

In this analysis, recognition of disfluency is judged to have been successful where subjects gave a judgment of "4" or "5". Word recognition was judged to be successful where subjects identified the correct word or a closely related word (eg "want" is taken as a correct recognition of "wanted", "was" is accepted for "were").

Using these criteria, the gate numbers at which recognition of disfluencies and words following the disfluent interruption point occurred and where the acoustic onset of these words were placed are compared. So for each disfluent utterance there are three points of interest: the gate in which the word following the interruption begins, the point at which the word is recognised and the gate at which the disfluency is recognised.

A total of 43 subjects each gave judgements on 7 of the 28 disfluent utterances, giving a total of 301 cells. Disfluency was recognised successfully in 257 (85.4%) cases. The word following the interruption was recognised in 191 (63.5%) cases.

The following results are illustrated in Fig. 1.

Disfluency recognition preceded word recognition in 66.5% (193) of 290 cases (the fluency judgements for one test item are disregarded in this comparison as the results were obscured by a misunderstanding of the relevant instructions). This result showed that, overall, subjects recognised that the utterance was disfluent before they had recognised the word following the interruption. A matched t-test was performed using only those cells where both disfluency and the crucial word were recognised (N=181) and the result was highly significant (t=-9.71, df=180, p<0.0001).

Word and disfluency recognition occurred at the same gate in 14.1% of cases and in 13.4% neither were recognised by the offset of the second word.

Word recognition preceded disfluency detection in only 5.9% (17) of cases.

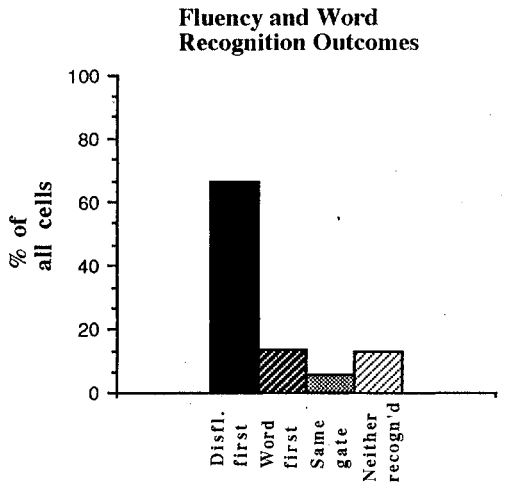

Figure 1.

Further t-tests examined the relationship between word onset and disfluency and word recognition. In 20% of all cases (but only to a significant degree in 4 items) subjects were able to detect disfluency before the *onset* of the word following the interruption (ie where there was an extended pause or where an mid-word interruption was detected): this led to no significant difference being found overall between the word onset point and the point of disfluency recognition (t=-1.59, df=180, p=0.1132), the mean difference between word onset gate and the gate at which disfluency was recognised being -0.54 (disfluency recognition following word onset). Since word recognition never occurred prior to the onset of the word, and occurred an average of 3.9 gates later, the difference was significant (t=-18.16, df=180, p<0.0001).

In fluent utterances, it was observed that non-recognition of a word had no significant effect on fluency judgements. Subjects were able to correctly judge that the utterance was still fluent even though they had not yet recognised the word that they were trying to identify.

2.6 Conclusion

In a significant number of cases, subjects were able to detect disfluency before they could recognise the word following the interruption. The non-recognition of words did not appear to affect fluency judgments: in fluent utterances, subjects were able to correctly judge fluency, despite not yet recognising the current word.

In a few cases disfluency was detected prior to the onset of the crucial word. The two main causes of this result are clear mid-word interruptions (eg "Ab- Aberdeen") and extended pauses.

3 DISCUSSION

The results suggest that listeners are able to recognise disfluency in an utterance on grounds other than lexical or syntactic.

A previous experiment with low-pass filtered speech using the same materials found that listeners were able to identify speech as disfluent without access to segmental information after the interruption point using only prosodic cues [12]. Prosodic information has been shown to be useful in processing fluent speech: Martin ([13], [14]) and Buxton ([4]) show that listeners make use of rhythmic expectancy in processing fluent speech; Darwin ([6]) shows that listeners pay attention to prosodic continuity in speech even to the extent that this information may overrride syntactic and semantic information.

It thus seems likely that listeners make use of expectations of prosodic continuity in processing speech with disfluencies and that prosodic information plays a primary rôle in resolving the processing problems presented by disfluent speech.

Work is currently under way to examine in detail the acoustic cues available to listeners at the recognition points of the disfluencies used in this experiment. In addition, another perception experiment using low-pass filtering and 35msec gates will determine how soon prosodic information alone provides enough information for the detection of disfluency.

Acknowledgements

The first author was supported by award number 87310722 from the UK Science and Engineering Research Council.

References

[1] E.G. Bard, R.C. 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(5):395–408, 1988.

[2] 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 Association for Computational Linguistics*. Association for Computational Linguistics, 1992.

[3] E.R. Blackmer and J.L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194, 1991.

[4] H. Buxton. Temporal predictability in the perception of english speech. In *Prosody: Models and Measurements*, volume 14 of *Springer Series in Language and Communication*. Springer-Verlag, Berlin, 1983.

[5] A. Cutler. Speakers' conceptions of the function of prosody. In *Prosody: Models and Measurements*. Springer-Verlag, Berlin, 1983.

[6] C.J. Darwin. On the dynamic use of prosody in speech perception. In A. Cohen and S.G. Nooteboom, editors, *Structure and Process in Speech Perception*, pages 178–193. Springer-Verlag, Berlin, 1975.

[7] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128. Association for Computational Linguistics, 1983.

[8] P. Howell and K. Young. The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology*, 43A(3), 1991.

[9] W.J.M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[10] W.J.M. Levelt and A. Cutler. Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–217, 1983.

[11] R.J. Lickley, E.G. Bard, and R.C. Shillcock. Understanding disfluent speech: is there an editing signal? In *Proceedings of the ICPhS*, volume 4, pages 98–101, Aix-en-Provence, France, August 1991. International Congress of Phonetic Sciences.

[12] R.J. Lickley, R.C. Shillcock, and E.G. Bard. Processing disfluent speech: How and when are disfluencies found? In *Proceedings of Eurospeech 91*, volume 3, pages 1499–1502, Genova, Italy, September 1991. 2nd European Conference on Speech Communication and Technology.

[13] J.G. Martin. Rhythmic (hierarchical) versus serial structure in speech and other behaviour. *Psychological Review*, 79(6):487–509, 1972.

[14] J.G. Martin. Rhythmic expectancy in continuous speech perception. *Communication and Cybernetics*, 11, 1975. In Cohen and Nooteboom (eds).