UNDERSTANDING DISFLUENT SPEECH: IS THERE AN EDITING SIGNAL?

R. J. Lickley, R. C. Shillcock and E. G. Bard.

Dept. of Linguistics and Centre for Speech Technology Research,
University of Edinburgh, Scotland

ABSTRACT

The problems posed by the frequent occurrence of disfluency in normal speech are important both for psycholinguistic and computational models of speech understanding. The most basic of these problems is determining when disfluency has occurred. Hindle [1] makes use of a phonetic 'editing signal' which marks the end of the material to be ignored and indicates the onset of the repair. This paper presents the results of gating experiments on spontaneous speech which show that only a minority of disfluencies can be detected by the point where this signal is claimed to occur, but that nearly all are obvious to listeners within the first word of the repair.

1. INTRODUCTION

Unlike written or read language, spontaneous speech is characterised by numerous disfluencies. For the purposes if this discussion, disfluency will be understood to consist of two main types: repetitions (Example 1) and false starts (Example 2). Both may be of lengths varying from less than a syllable to several words. Other hesitation phenomena - silent and filled pauses and lexical fillers - will not be discussed.

Example 1: Repetition:

'And you'd re- you'd really need about eight ...'

Example 2: False Start:

'Because although the bell the rules say that ...'

It is all to easy to miss disfluencies when transcribing spontaneous speech verbatim, and all too difficult to believe that so many occurred when perusing a correct transcription because we appear to notice very few of them as they occur.

One of the factors which may facilitate the processing of disfluent speech could be the presence of cues in the speech stream prior to the break in fluency which prepare listeners for a break. Don Hindle [1] makes use of this idea in his algorithm for parsing speech with disfluencies:

> *'Two features are essential to the self-correction system: 1) every self-correction site [...] is marked by a phonetically identifiable signal placed at the right edge of the expunction site ...'*
> ([1] p128)

Hindle's editing system depends crucially on the presence of this editing signal (see Labov [2]), defined as [1]. The system takes as input a transcription in standard orthography of conversational speech which has editing signals inserted by the transcriber, *when noted*, at the point of interruption.

The experiments described in this paper are designed to establish the location of the editing signal to a first approximation. They use materials from a sample of repetitions and false starts drawn from and representative of those in a corpus of studio-recorded spontaneous conversational English. The first experiment establishes that listeners are able to recognise that an utterance is disfluent by the offset of the first word following a disfluent interruption. The second experiment addresses Hindle's supposition that an editing signal *'placed at the right edge of the expunction site'* (ie immediately following the section of speech that is to be ignored and prior to the onset of the continuation) indicates to the listener that a disfluency is present. It is found that the majority of disfluencies are not detectable at this point in the utterance. The conclusion is reached that, if an editing signal is present in disfluent speech it is not as a discrete phonetic signal, but rather a feature of the prosodic disruption that takes place.

2. EXPERIMENT ONE

2.1. Introduction

This experiment was designed to test the hypothesis that disfluency can be recognised by the offset of the word following the interruption point.

2.2. Materials

From a corpus of spontaneous speech, recorded digitally in a studio, 30 *spontaneous disfluent* utterances were selected, each containing a token of one of a set of types of disfluency, to be used as test items. The types of disfluency and the numbers of each type used were representative of the distribution of types of disfluency identified in the corpus by the first author. Test items were divided equally among the six speakers whose conversations make up the corpus.

Next, another 30 utterances were chosen from the corpus to provide *spontaneous fluent* controls for the disfluent items. These items were selected to match the disfluent utterances for structure, length and prosody as far as possible.

To provide controls better matched in structure to the spontaneous disfluent utterances, each such item was edited using ILS to remove the disfluency and leave, without interruption, the fluent parts of the utterance. Each of the original speakers then heard the doctored versions of his or her utterances and was asked to produce 6 fluent imitations of each. The speakers' responses were recorded under the same conditions as in the recording of the original conversations. For each item, the most accurate of the imitated versions was selected to be the control for that item, accuracy being defined as closest matching in terms of rate and rhythm of production.

Examples of the resulting test materials are given below.

Example 3:

Spontaneous Disfluent:

'... it's quite obvious he's he's on something ...'

Rehearsed "Disfluent":

'... it's quite obvious he's on something ...'

Spontaneous and Rehearsed Fluent:

'... we know that it's not going to ...'

All the utterances to be used were sampled on ILS on MASSCOMP through a 8kHz filter at 20kHz, together with up to 10 seconds of the conversation which occurred prior to the test utterance, which provided some discourse orientation. The onset of each word in each item was determined from a combination of auditory information and time-amplitude waveform. Each item was then *gated* at word boundaries so that the first stimulus for an item ran from its onset to the end of its first word (*it's*), the second from its onset to the end of its second word (*it's quite*), the third to the end of its third word (*it's quite obvious*) and so on.

The test materials were divided into two complementary sets of sixty utterances so that neither of the two sets of subjects heard both the spontaneous and the rehearsed versions of any utterance. Each set of 60 items was blocked by speaker and recorded on a separate test tape.

2.3. Subjects and Procedure

Twenty students and staff members of the University of Edinburgh served as subjects, 10 per group. All were native speakers of English familiar with the range of accents represented in the

experimental materials and all reported having normal hearing.

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

Subjects were given adequate time to familiarise themselves with each speaker's voice and all utterances were presented with about ten seconds of the dialogue prior to the utterance.

There were two tasks in the experiment: word recognition and disfluency recognition. For the word recognition task, subjects were asked to write down after each gated presentation what they thought the latest word presented was and to make any amendments required to previous words in the appropriate part of the answer sheet. For the disfluency recognition task, subjects were asked to make a judgement on a 1-5 scale about whether they considered that the utterance was fluent at the current word gate. A score of 1 indicated that the subject considered that the utterance was fluent, a score of 5 indicated detection of disfluency and intervening scores indicated uncertainty.

2.4. Results

In this analysis, only the 1-5 scores for the crucial point in the disfluent utterances (the first word of the restart) and the equivalent points in the control utterances are examined.

Subjects were able to give fluency judgements with considerable confidence. For disfluent utterances, they gave average scores of between 4 and 5 in the majority of cases (max = 50, min = 17, mean = 40.05); the controls received average scores of 1 or just over 1 (min = 10, max = 48, mean = 12.39, for all controls).

The differences between fluency judgements for critical points in disfluent utterances and the equivalent points in the controls were found to be significant (Friedman statistic by subjects = 38.2, $df = 3, p < .001$; by materials = 50.91, $df = 3, p < .001$).

There were 2 cases out of the total of 30 disfluencies where the total score for the disfluency judgement was lower than 30, indicating that on average subjects thought that the utterance might still be fluent. These scores were examined individually in Wilcoxon signed rank tests, comparing them with the scores for their fluent controls: there was still found to be a significant difference between the sets of scores, the scores for the disfluent items being higher than for their fluent controls (first case: $n=6$, $W=0$, $p<.025$; second case: $n=7$, $W=0$, $p<.01$).

2.5. Discussion

The subjects gave high scores of between 4 and 5 in the majority of cases where disfluency had occurred and low scores of between 1 and 2 where there was no disfluency, thus supporting the hypothesis that disfluency can be recognised by the offset of the first word after disfluent interruption.

3. EXPERIMENT TWO

3.1. Introduction

This experiment was designed to test the hypothesis that an editing signal at the interruption point prior to the continuation enables listeners to detect disfluency.

3.2. Materials

The materials used in this experiment were identical to those used in the first.

3.3. Subjects and Procedure

There were 20 subjects, as in the first experiment.

The procedure was the same as that in the first experiment except that the disfluency recognition task differed: subjects were asked to use the 1-5 scale to say whether they thought that, on the basis of what they had heard, the utterance would *continue* fluently or disfluently. Thorough explanations and practice sessions preceded the experiment.

3.4. Results

In this analysis, the critical point in the utterance is the word-gate prior to the restart.

Subjects showed less confidence in their fluency judgements than in the first experiment. They gave average scores of between 2 and 3 for the critical point in disfluent utterances (max = 3.7, min = 1.3, mean = 2.55); the average scores for the equivalent point in the controls were of 1 or just over 1 in most cases (min = 1.0, max = 3.7, mean = 1.9, for all controls).

The differences between fluency judgements for critical points in disfluent utterances and the equivalent points in the controls were found to be significant (Friedman statistic by subjects = 34.62, $df = 3, p < .001$; by materials = 21.77, $df = 3, p < .001$).

To examine the results for individual test items, Wilcoxon signed rank tests were performed, comparing scores for the spontaneous disfluent condition with those for the spontaneous fluent condition. The results of these tests show that the scores for the disfluent condition were significantly higher than those for the fluent condition in only 12 of the 30 cases ($p<.05$), the difference in scores was insignificant in 15 cases and the difference was significantly higher for the fluent condition in 3 cases.

3.5. Discussion

The results show that the hypothesis is only supported by a minority, 12, of the 30 test items. Of these 12, only 9 have average scores of 3 or over and the maximum is 3.7, which should indicate that subjects had a slight feeling that disfluency was about to occur.

A reexamination of the materials to search for any phonetic cues which may have caused higher scores reveals that the 12 test items for which the total scores were 30 or over fall into one of two main categories: words which are interrupted suddenly (incomplete words); words which are lengthened and/or followed by a pause and/or creaky offset or an inbreath. The majority of the other test items consist of complete words with no pause before the continuation.

The analyses suggest that listeners made use of cut-offs and hesitation phenomena, where they were present, in detecting oncoming repairs, but in the majority of cases, where such cues were not present, they were unable to detect imminent disfluency.

4. CONCLUSION

The experiments reported in this paper show that disfluency can usually be detected by the end of the first word following the interruption and do not support the hypothesis that listeners perceive and make use of a phonetically identifiable editing signal placed immediately prior to the onset of the continuation. Subjects only indicated that they detected oncoming repairs in a minority of cases. In the majority of cases, they appeared to make use of cues within the first word of the repair.

Further experiments are under way to determine more precisely where listeners can detect disfluency and to examine the contribution of prosodic cues to the perception of disfluency. It is suggested that rhythmic and intonational information plays a vital role in alerting listeners to the presence of disfluency, rather than a discrete phonetic editing signal.

7. REFERENCES

[1] HINDLE, D. (1983), "Deterministic Parsing of Syntactic Non-Fluencies", *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*

[2] LABOV, W. (1966), "On the Grammaticality of Everyday Speech", *Paper presented at the Annual Meeting of the Linguistic Society of America.*