JUNCTURE CUES TO DISFLUENCY

R. J. Lickley

Department of Linguistics
and Human Communication Research Centre,
University of Edinburgh, UK

ABSTRACT

This paper describes properties of normal disfluent speech which help listeners to distinguish disfluent from fluent strings of speech. It focusses on juncture phenomena in cases where there is no clear silent pause at the interruption point. Recent attempts to define acoustically identifiable features of speech which can be seen as reliable indicators of disfluency have produced several suggestions. But studies of silent pause, (pre-)pausal lengthening, glottalisation and measurements of F0 have all failed to provide any reliable means of distinguishing fluent from disfluent continuations.

This paper introduces into the discussion a phonological feature of speech which has been overlooked in previous work and which could prove to be a reliable indicator of disfluency, especially in mid-clause disfluencies where no pause is present at the interruption. In normal fluent continuous speech, words are not usually separated by silent pause into discrete units, but have their boundaries obscured or linked by processes like assimilation, liaison, elision and so on. The hypothesis examined by the present study is that such juncture phenomena are blocked by disfluency. Evidence from perceptual experiments suggests that this phenomenon may be used by human listeners in early detection of disfluency.

1. INTRODUCTION

Normal spontaneous speech is characterised by frequent occurrences of repetitions (*"go to the – to the edge of the mountain"*) and false starts of various types (*"you've got one to – have you got a lake on the left?"*). Such disfluencies present problems for both computational and psychological models of speech understanding.

Most computational models have taken a text-based approach, choosing to view the problem as one to be resolved primarily by parsing or pattern-matching on words which have already been recognised, using orthographic transcriptions (e.g. [3, 5]). Hindle [10] suggests that a phonetically identifiable editing signal accompanies every disfluency. When such a signal is seen by the processor, it triggers a series of copy editors which find matching orthographic strings or matching syntactic categories on either side of the editing signal and expunges them from further analysis. However, no clear definition of the editing signal has ever been proposed, and, like others, the model depends on successful prior recognition of words.

Nakatani and Hirschberg [19, 20] propose a model of repair detection which takes into account speech-based cues, but still relies heavily on textual cues, such as the presence of fragments and lexical and part-of-speech information.

Research on human perception, however, suggests that prosodic, rather than syntactic or semantic information is responsible for our impression of fluency in speech. Darwin [7] found that subjects in a dichotic shadowing task would temporarily follow prosodically fluent speech when the semantic context had become anomalous, even when they had the opportunity to follow the semantically fluent message. Other studies have shown that listeners can avoid or quickly resolve potential syntactic ambiguities by means of prosodic information available early in the sentence (e.g. [2, 17, 18]), thus, for example, overriding the garden path effects which a text-based processor might suffer and saving the processor from complex parsing problems. An additional finding in human speech processing, which is not taken to account in the text-based approach is that words are not all recognised in the order in which they are spoken: many words can not be identified until after following words have been heard (e.g. [1, 6]). This makes it uncertain that at a given point in an utterance a listener will have a clear enough picture of the context to judge whether a syntactic anomaly is present.

These psycholinguistic studies only examine the processing of fluent speech: the aim of the present paper is to propose that similar features in *disfluent* speech may allow potential parsing problems faced by text-based processors to be avoided similarly early.

In the following sections we discuss briefly some psycholinguistic experiments which addressed the questions of how and when listeners can detect disfluency. We then discuss acoustic and prosodic cues that have been proposed by other researchers and propose our own interpretation, suggesting

that the absence of the juncture phenomena found in normal fluent speech can provide useful cues. The cues proposed are then looked at in the light of the results of the experiments and we attempt to identify which of the cues could have been of use to listeners when they detected disfluencies.

2. EARLY DETECTION OF DISFLUENCY: EXPERIMENTAL EVIDENCE

Few psycholinguistic studies have specifically examined disfluent speech. But previous work by this author has provided some evidence of how soon sufficient information is available for listeners to be able to detect disfluency. Experiments reported in [14, 13, 15, 16] show that listeners are able to detect disfluency very soon after the interruption point and without necessarily having recourse to lexical or syntactic information. In word-level gating experiments (where utterances were presented in chunks increasing by one word on each presentation), it was established that listeners could usually detect disfluency when they heard the first word after the interruption. In a subsequent 35msec gating experiment, it was found that disfluencies could often be detected very soon after the onset of the repair and even before the first word of the repair could be recognised. So something other than lexical or syntactic information seemed to tell the listener that the utterance was disfluent.

Two further experiments using the same materials presented the stimuli low-pass filtered at such a level that segmental information was not perceptible, but that voicing, F_0, and durational information remained. In the first of these tests, subjects demonstrated that they could discriminate between disfluent utterances low-pass filtered from the interruption point and matched fluent utterances filtered from a similar point. In the second experiment, using 35msec gating in addition to low-pass filtering, it was shown that subjects could still recognise disfluency within one word of the interruption point, even with segmental information effectively removed.

This work suggests not only that disfluency can be detected very soon after the interruption, but also that listeners do not have to rely on syntactic information to the extent suggested by computational models.

3. POSSIBLE CUES

So what cues did subjects make use of in their successful early detection of disfluency? Previous studies have suggested several types of cue, located around the interruption in three distinct domains: in the reparandum (the part of the utterance that is in effect replaced by the repair), in the editing phase (from the end of the reparandum to the start of the repair), and in the repair itself, where speech resumes after the interruption. Glottalisation at the end of the reparandum has been suggested as a possible cue [19, 20, 3], particularly in vowel-final fragments. Silent pause at the interruption point is seen as another indicator (see, for example [4], on the problem of how to define a pause), though (pre)pausal lengthening may provide a cue in a similar way, without any actual silence occurring [8]. Filled pauses (*um, uh*) and lexical fillers (*well, I mean*) are also seen as potential markers of oncoming repair. Prosodic features of the repair phase that have been suggested as cues are limited to the use of contrastive stress in certain conditions [12], which can not be generalised to the majority of disfluencies. Other studies have compared F_0 across repairs to matches selected on syntactic grounds [3], have been limited in scope to single word repetitions [21] or have taken measurements from accented syllables around the interruption [20], rather than the immediate context, which is our focus here.

One aspect of normal fluent speech which most other studies seem to have overlooked in their search for cues to disfluency is that not only are words *not* separated by pauses into discrete units, but that their boundaries are usually linked or obscured by phonological processes like coarticulation, assimilation, liaison, degemination, elision and sandhi phenomena. These links are so smooth in continuous speech that it is often impossible to segment the speech into words on the basis of acoustic information. It is well known that syntactic and prosodic boundaries often "block" such linking, but the likelihood of such blocking depends greatly on speech style and speech rate (e.g. [11]): in the faster, casual speech that occurs in spontaneous conversation, such boundaries are not respected by all types of linking, all the time.

4. ANALYSIS

Of the possible cues mentioned above, we exclude filled pauses and lexical fillers from our analysis, as they did not occur in our experimental materials. In any case, our speech corpus [13], like others, showed that such fillers do not typically mark the interruption in repetitions and false starts. Only 6.54% of 1146 repairs were accompanied by filled pauses and only 5.24% were accompanied by lexical fillers: these figures represent 10.64% of filled pauses (N=528) and 11.36% of lexical fillers (N=528).

4.1. Method

Materials analysed were 30 disfluent utterances taken from a corpus of spontaneous British English dialogues described in [13], 30 fluent utterances from the same corpus, chosen as matches for the disfluent items on the basis of syntactic and prosodic similarity, and 30 rehearsed utterances, which were fluent versions of the disfluent utterances, produced by the same speakers. All had been used as stimuli in the experiments described in section 2, above. All utterances were sampled at 16kHz and analysed on a Sun Sparcstation, using the Entropic ESPS/Xwaves+ software.

Pauses were defined as periods of no vocal activity in the middle of utterances, which did not merely coincide with consonantal stop closure or glottal closure before a vowel-

initial word. They were measured by marking the points on the waveform where the pause began and the point where the onset of the resumption of the signal could be detected by visual and auditory examination. Where the restart began with a stop consonant, a 50msec closure phase was taken into account. F_0 measurements were taken on the word prior to and the word following the interruption point in disfluent items and at equivalent points in the fluent controls. Other observations were made by close examination of the waveform, pitchtrack and spectrogram and by playback of the signal.

Sample sizes of the phenomena within the data set used in the experiments are too small to treat with statistical tests but the interest in the study is that proposed cues can be compared with responses in an experiment where subjects were asked to judge whether or not repeats of false starts were present in the stimuli. To this end, histograms illustrating mean fluency judgements from the 35msec gating experiment described above were aligned on the screen with the visual representations of the signal so that increases in disfluency detection rates could be matched with the acoustic and prosodic features being examined, to within 35msec of their occurrence.

4.2. Results

Glottalisation. Reparanda were classed as "glottalised" where they ended in a vocal section with a deceleration in the rate of vibration in the vocal cords or a glottal stop.

In our sample, 9 of the 30 disfluent items had glottalised reparandum offset including 2 of the 5 vocal-final fragment-final reparanda (there were 9 fragment-final reparanda in all). In all but one cases, the glottalisation was followed by immediate repair, with no silent pause. None of the control utterances displayed glottalisation at the word boundaries selected as matching the disfluent boundaries. It is not clear whether a distinction can be made between interruption glottalisation and other types of glottalisation (or laryngealisation) which occur in fluent speech, although examination of some examples in our corpus suggested that the former shows tenseness of the vocal cords and a rapid slow-down to stop of glottal vibration, where the latter shows slow but regular vibrations and laxness characteristic of creaky voice.

Comparison of mean disfluency judgements with episodes of glottalisation at the same points in time did not reveal an immediate effect. Subjects preferred to wait until the beginning of the repair. If there was any effect, it may have been delayed and combined with other cues.

Pauses. Sixteen of the 30 disfluent stimuli contained silent pauses of between 34msec and 1134msec (mean=278msec). In the controls, one pause was found at a matched point in a spontaneous fluent case, but none in the closer-matched rehearsed versions. That pauses were unexpected at the interruption points was confirmed by looking at the phonological phrase structure of the disfluent items, assuming fluent continuations. In all but one case, where a clause boundary coincided with the interruption, no pause would have been expected in normal production of the utterance.

As with other stimuli, in most cases where pause was present, subjects only showed that they were certain that disfluency was present once they heard the onset of the repair. One exception to this is where a loud inhalation is audible in the pause, but even in that case, the judgements of disfluency do not peak until the repair begins.

Pitch. Comparisons between F_0 measurements before and after the interruption point in disfluent stimuli and at equivalent points in fluent stimuli revealed little of statistical interest. A significant difference was found for F_0 differences from the word before to the word after the interrupt between disfluent and spontaneous fluent utterances ($t = 3.69, df = 29, P < 0.001$) but not between the disfluent stimuli and their rehearsed controls.

Structural analysis of the false starts in the sample showed that in 9 of the 11 cases, the restart was sentence-initial and had typical sentence-initial prosody (high F_0 and intensity). But these restarts were not always higher in pitch than the preceding peak (in the reparandum), because in most cases the preceding peak was itself sentence-initial. However, the importance of this prosody from the perceptual point of view is not necessarily in how it compares to the preceding prosody, but in the fact that it is unexpected: the intensity and F_0 associated with sentence onset is rarely found within fluent sentences, but its perceptual distinction from other types of prosodic peaks such as heavy emphasis and contrastive stress is yet to be tested.

For other types of disfluency no reliable trends could be identified from F_0 measurements. (See Shriberg's thesis [21] for F_0 analysis of a large number of single-word repetitions.)

Juncture. A large proportion of the disfluent stimuli used in the experiments had no silent pause at the interruption. These stimuli were examined to determine whether disfluency had blocked the phonological linking expected in fluent speech. To address the question, waveforms and spectrograms were examined for 13 stimuli for which it was possible to hypothesise a fluent continuation and which had no more than 50msec of silence at the interruption point. The boundaries between the offset of the reparandum and the onset of the repair were compared to hypothesised fluent boundaries between the same phonetic segments. For example, where the words on either side of the interruption were "we we", it was hypothesised that a fluent boundary would show smooth formant transitions from [i:] to [w], with steady voicing; where the boundary was between "the" and "over", it was hypothesised that "the" would end in [i:] and link smoothly to [ou] via [j].

Twelve of the thirteen stimuli examined had boundaries

which differed from the hypothesised fluent boundary. A common factor was that the repair onset usually commenced as if it was being produced in isolation, rather than as if it was preceded by a contiguous phonetic context. Repairs with voiced onsets commenced with glottal stops: *"el- eligible"* contained a glottal stop where a fluent progression from [l] to [e] would show smooth formant transitions; *"was was"* contained a glottal stop and a glottalised glide into the vowel, where a fluent link might contain assimilated lip-rounding in the fricative. Other cases involved the offset of the repair: in *"over the over the"*, "the" at the interruption was pronounced with schwa, as if a consonant was expected, rather than the [i:] hypothesised; in *"ab- aberdeen"* the interrupted [b] is not audibly released.

5. DISCUSSION

Juncture phenonema which occur between words in fluent speech are usually absent at the interruption point in disfluent utterances. Our experimental evidence suggests that the absence of such linking may be perceived by listeners and serve as cues to the early detection of disfluency. Such cues are comparable to those available to listeners in avoiding syntactic ambiguities in fluent speech [2, 17, 18]. It is, of course, likely that such cues interact with many other cues at different linguistic levels, but these cues seem to enable listeners to decide that an utterance is disfluent before having access to complete lexical and syntactic information.

6. 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. C.M. Beach. The interpretation of prosodic patterns at points of syntactic structure ambiguity: evidence for cue trading relations. *Journal of Memory and Language*, 30(6):644–663, 1991.
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 Association for Computational Linguistics*, 1992.
4. A. Butcher. *Aspects of the speech pause: phonetic correlates and communicative functions.* PhD thesis, Christian-Albrechts-Universität Kiel, 1981.
5. J.G. Carbonell and P.J. Hayes. Recovery strategies for parsing extragrammatical language. *American Journal of Computational Linguistics*, 9(3-4):123–146, 1983.
6. C.M. Connine, D. Blasko, and M. Hall. Effects of subsequent sentence context in auditory word recognition – temporal and linguistic constraints. *Journal of Memory and Language*, 30(2):81–94, 1991.
7. 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.
8. D. Duez. Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research*, 22(1):21–39, 1993.
9. F. Grosjean and J.P. Gee. Prosodic structure and spoken word recognition. *Cognition*, 25:135–156, 1987.
10. D. Hindle. Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, 1983.
11. R. Lass. *Phonology: an introduction to basic concepts.* Cambridge University Press, 1984.
12. W.J.M. Levelt and A. Cutler. Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–217, 1983.
13. R.J. Lickley. *Detecting Disfluency in Spontaneous Speech.* PhD thesis, University of Edinburgh, 1994.
14. R.J. Lickley and E.G. Bard. Processing disfluent speech: Recognising disfluency before lexical access. In *Proceedings of The ICSLP*, pages 935–938, Banff, Alberta, Canada, October 1992.
15. R.J. Lickley, E.G. Bard, and Shillcock R.C. 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.
16. R.J. Lickley, Shillcock R.C., 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.
17. W.D. Marslen-Wilson, L.K. Tyler, P. Warren, P. Grenier, and C.S. Lee. Prosodic effects in minimal attachment. *Quarterly Journal Of Experimental Psychology Section A – Human Experimental Psychology*, 1:73–87, 1992.
18. H N Nagel, L P Shapiro, and R Nawy. Prosody and the processing of filler-gap sentences. *Journal of Psycholinguistic Research*, 23(6):473–485, 1994.
19. C Nakatani and J Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 46–53, 1993.
20. 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.
21. E.E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, University of California at Berkeley, 1994.