Detecting Disfluency in Spontaneous Speech

Robin J. Lickley

A thesis submitted in fulfilment of the requirements
for the degree of Ph.D.
to the
University of Edinburgh
1994

Abstract

This thesis reports a study of the perception of disfluency in spontaneous, conversational speech.

Disfluent speech presents problems for both computational and psycholinguistic models of speech processing. The surface strings produced when speech is interrupted by disfluency require complex editing processes from computational models in order to produce well-formed strings for parsers. There is little empirical evidence about how the human speech processing mechanism deals with disfluencies, but our everyday experience of listening to speech suggests that we can deal with disfluencies very smoothly and efficiently. One of the first problems for a speech processor is to detect that disfluency has occurred. No reliable acoustic or prosodic cues have been identified which signal the presence of a discontinuity. In this thesis, the main aims are address this problem by first establishing detection points for a set of disfluent utterances and then finding out what acoustic and prosodic cues are available at these points.

The main part of the study consists of a series of 5 perceptual experiments, followed by acoustic and prosoodic analyses. The first 3 experiments establish detection points for disfluencies and relate these points to recognition points of the words in the vicinity of the interruption. The last 2 experiments examine the rôle of prosodic information in detecting disfluency., first over whole utterances and then focussing in on the region of the interruption. The acoustic and prosodic analyses of the experimental stimuli match responses indicating disfluency detection to events in the speech signal which might act as cues.

The results of the first 3 experiments show that disfluency can be recognised very early, usually within the first word after the interruption point. Importantly, it is also shown that the detection of disfluency can be achieved before the word is

recognised – non-syntactic information is used. The last two experiments confirm that prosodic information can be used to distinguish fluent from disfluent utterances. The acoustic and prosodic analyses suggest that a combination of cues can be of use. In the absence of "ungrammatical pause" and broken-off words, a break in the signal is signalled by the absence of phonological linking between the words on either side of the interruption. It may be possible to identify other cues in future studies with larger data sets.

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

January 1994

Acknowledgements

It would be impossible to mention by name all the people who have contributed to this thesis in terms of academic, technical and moral support over the past few years: a large number of colleagues, friends and squash partners have played a part and I am sincerely grateful to every one of them.

My PhD Committee of Ellen Bard, Richard Shillcock and Steve Isard have always been generous with time and advice and have had a great influence on the work. Ellen Bard is everything anyone could wish for in a supervisor: I have benefited a lot from her insightful guidance and inspiration as well as her encouragement and patience.

Many other people deserve acknowledgement on the academic front. Among these, I am especially grateful to Liz Shriberg, for useful and stimulating email discussions of our topic and for showing me that I wasn't the only person in the world "doing disfluency", and to my Edinburgh psycholinguistics colleague, Louise Kelly, for helpful discussions of statistics and word recognition and plentiful supplies of biscuits and drinking chocolate.

I gratefully acknowledge the financial support of the Science and Engineering Research Council (Award number 87310722). The Centre for Speech Technology Research was also generous in employing me at the end of my SERC funding and allowing me continued use of computing facilities.

Many thanks are also due to the computing and technical staff both in the Department of Linguistics and the CSTR.

Finally, my heartfelt thanks go to my parents for being so supportive and encouraging throughout my academic career and to Nucy for her support, patience and fortitude over the past few years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Um no I uh I don't know I find the Ro- Romans anyway very I don't know there's a [pse] they they live through their senses more than the British do in a way and uh I find I f- f- I think that's fine for a holiday but for any length I think it would start getting [pse] uh I'd just start feeling an outsider you know. No I need a I need a certain depth of community which [pse] which I didn't see even in Siena which is a small [pse] place um there's still the same um um [dhe] there's still that atmosphere
>
> ...
>
> but nevertheless it's still [pse] it's [?I] it I don't know as an outsider it can make you feel my my point was really that you feel it makes you feel uh displaced or disorientated being [pse] with with priorities and values like that when they're not your own you know?

Everyday conversational speech is characterised by the highly frequent occurrence of **disfluency**. Filled pauses (*um, uh*), repetitions (*they they*), false starts (*my point was really that you feel it makes you feel ...*) and other phenomena are the norm, rather than exceptions. Transcriptions of spontaneous speech often provoke surprise and disbelief in the reader, because the textual discontinuities are so evident when presented in written form[1]. And yet when we hear

[1]The above excerpt is from a casual conversation recorded as part of the corpus for the

such speech, we can apparently follow and understand it with the minimum of trouble.

This thesis is a study of the perception of disfluency in normal speech. To clarify the scope of the research and to forestall any mistaken preconceptions, it is important to declare what it is *not* about, first of all. The study is not directly concerned with stuttering or other pathological speech behaviours: the informants in our corpus are all "normal" speakers. Nor is it concerned directly with "speech errors" or "tongue slips", which have attracted much attention in psycholinguistics for the information they can purvey about speech production (Fromkin, 1980; Cutler, 1982): the "classic" cases of tongue slips often pass uncorrected and are produced quite fluently; they are also relatively rare. The only occasions where tongue slips *do* play a part in this investigation is when they are immediately corrected, or repaired by the speaker: when this occurs, the error becomes part of a disfluency. Finally, the thesis is not concerned with personality and emotional characteristics of disfluent speech (Mahl, 1957; Kasl & Mahl, 1965) nor with perceived character traits of disfluent speakers (Miller & Hegwill, 1964): no attempt was made in the present study to elicit disfluencies by putting informants under stress or by distraction (Yngve, 1973) – our corpus of speech was collected in the setting of a casual conversations in a relaxed situation.

The spontaneous speech phenomena which are of interest for this investigation are the types of normally-occurring ungrammatical repetitions, false starts and disfluent pauses which are illustrated in the above excerpt. The nature of the perceptual problems to be examined concern the resolution of the trouble caused by disfluencies in on-line processing.

Computational models of speech processing which address the problem of extracting the intended message from "ill-formed input" usually rely primarily on syntactic information. They also make the psycholinguistically fanciful assumption that all words are segmented and labelled prior to syntactic processing. For such models, special algorithms are activated when an initial parse has failed, which detect certain patterns which can occur in disfluency and attempt to progressively remove sections of the sentences in order to reduce the

---

present study: the speaker is a "normal" speaker of standard English

input to a parsable sentence. This turns the problem of disfluency resolution into a somewhat complex and computationally cumbersome operation, which is not guaranteed to succeed in any case (the problems with computational approaches are further discussed in section 2.2.2).

The present research is particularly concerned with the problem of how disfluency can be detected in on-line processing. Research in speech production (e.g. Levelt, 1983; Blackmer & Mitton, 1991) suggests that speakers are very good at monitoring their own speech for errors and correcting them. But very little is known about how the *listener* copes with speech that has been repaired. Our everyday experience of listening to speech suggests that we are able to process disfluent speech very quickly and are usually unaware of the presence of discontinuities. But hardly any previous work has investigated human processing of disfluency. This thesis approaches the problem by asking two fundamental questions which have not been previously addressed:

1. **How soon** can the listener detect disfluency?

2. **What cues** can the listener use in detecting disfluency?

The questions are approached from two angles. First, a series of perceptual experiments find recognition points for both words and disfluencies and assess the ability of the listener to use prosodic information in detecting disfluency. Then, acoustic and prosodic analyses are applied to the stimuli, with reference to the detection points established in the experiments, in order to discover what information was present in the signal at the points where disfluency was detected.

1.1 Structure

Chapter Two provides the background for the investigation. It begins by surveying the literature regarding the typology of disfluent speech, and describing the types and terms to be used for the rest of the study. The second section looks at previous approaches to the processing of disfluent speech: most of the relevant work come from computational linguistics, rather than psycholinguistics, and makes assumptions about the input to the speech processing mechanism which

are incompatible with on-line processing of speech. Possible cues to disfluency are discussed in the third section: the literature specific to disfluency is fairly thin in this area but it is useful to look at some possible cues suggested by inference from the study of fluent speech as well as surveying the types of cue that *have* been proposed. The final section describes the experimental methodology used, introducing the *gating* paradigm and the technique of *low-pass filtering* of speech.

Chapter Three describes the recording, transcription and textual analysis of the corpus of spontaneous speech used in the rest of the thesis. The selection of stimuli for use in the experiments is explained.

Chapter Four is the first of three chapters which describe the experiments carried out to address the research questions. Experiments One and Two use word-level gating to find recognition points for disfluency to a first approximation. Experiment One tests the hypothesis that a signal prior to the onset of fluent speech after the interruption alerts the listener to the presence of disfluency. Experiment Two tests the hypothesis that disfluency can be detected within one word of the interruption. Both experiments allow hypotheses to be tested regarding the recognition of words in the immediate vicinity of the interruption.

Chapter Five describes an experiment with 35ms gates, which allows us to test the hypothesis that disfluency can be detected before the word after the interruption has been recognised. This experiment provides accurate "detection points" for disfluencies, which are used later in acoustic analysis. A control experiment checks for any artefactual effect of the dual task of word recognition and disfluency detection.

In **Chapter Six**, two experiments with low-pass filtered speech examine the use of prosodic information in detecting disfluency. Experiment Four presents subjects with whole utterances, the ends of which are low-pass filtered so that no segmental information is audible, and asks them to distinguish disfluent from fluent utterances. Experiment Five combines low-pass filtered speech with 35ms gating, to test the ability of listeners to detect disfluency soon after the interruption using only prosodic information.

Chapter Seven examines the acoustic and prosodic information in the signal at the crucial area in disfluent stimuli, to attempt to define what aspects of the signal subjects in the perception tests took to be cues to disfluency.

Chapter Eight summarises the findings of the experiments and acoustic and prosodic information, assesses the fulfilment of the aims of the thesis and suggests some possible directions for future research, both in experimentation and in analysis of the signal.

Chapter 2

The Background

This chapter is intended to provide the context for an understanding of the rest of the thesis. There are four main sections. In the first section the terminology and typology that has been used to describe disfluency is discussed and the terms and types to be used in the description of data in this thesis are defined. The second section discusses approaches to the processing of disfluent speech. No psycholinguistics approaches have really looked in detail at how people understand disfluent speech: some of what *has* been done is described in this section. The problem has been approached by more researchers in Computational Linguistics: the second part of section 2.2 describes some of the main CL approaches. Acoustic and prosodic information is likely to have an important part in the detection of disfluency: the third section discusses what sort of cues may be available, based first on a survey of some relevant work on the processing of *fluent* speech and then on the types of cues that other studies of disfluency have suggested. The fourth section discusses the choice of methodology for the experiments.

2.1 Terms, Types and Definitions

This section discusses the terminology to be used to describe the speech phenomena with which the thesis is concerned. The literature which provides input to a study of the phenomena in general is diverse, coming from fields such as speech

pathology, speech production, pragmatics, conversation analysis, discourse analysis, social psychology, computational linguistics, artificial intelligence as well as psycholinguistics. As a result of the variety of approaches, many different terms are found in the literature to describe what are often the same phenomena. In addition, many different systems of categorisation mean that different divisions are made in the data according to the research orientation of specific studies. This section will not be an attempt to unravel the potential confusion in this myriad of terms and definitions, but will describe some of the many approaches and then explain the selection of terms used in this study and define their rôles in the description of the data.

2.1.1 Disfluency

The phenomena which are referred to in this thesis by the generic term "disfluency" have fallen under several other headings in the literature. "Pause" (e.g. Duez, 1993), "hesitation" (e.g. Maclay and Osgood, 1959), "disturbance" (e.g. Kasl and Mahl, 1965) "fragmentation" (Allen & Guy, 1974), "hemming and hawing" (Hockett, 1958), "non-fluency" (Hindle, 1983), "speech management" (Allwood *et al.*, 1989),"discontinuity" (Taylor & Cameron, 1987), "repair" (e.g. Cutler 1983) and "self-repair" (e.g. Levelt, 1983, 1989) have all been used to refer to more or less the same set of phenomena. Little discussion of the choice among these terms has appeared in the literature, and only a brief discussion follows here, concentrating on the most frequently appearing terms, "pause", "hesitation", "(self-)repair" and "disfluency".

So why do we choose to use "disfluency" rather than "repair", "pause" or "hesitation"?

The terms "pause" and "hesitation" carry the implication that the speech signal is stopped for a period of time when speakers interrupt themselves. But sometimes speakers seem precipitous, rather than hesitant, in editing their speech: as Blackmer and Mitton (1991) show, and as is found in our own data, the time between a self-interruption and a restart is often zero. In addition, "pause" is widely used to refer to all temporal breaks in the speech signal, whether they be fluently-occurring juncture pauses (at major syntactic boundaries) or "ungrammatical"

silent or filled pauses which simply interrupt the speech flow momentarily before the speaker recontinues, often with no break in syntactic coherence. So, since "pause" and "hesitation" can be misleading, they are rejected as generic terms for use in this thesis.

"Repair" refers to the process whereby a speaker makes a change to their speech output. Examples of such change are error correction, word substitution, and qualification of something the speaker has just said or, as we shall see, was about to say. "Self-repair" is often used in order to distinguish between speaker-initiated change and other-initiated change in conversational interaction ("interactive repair", Couper-Kuhlen, 1992; Schegloff *et al.*, 1977): but where it is clear from the context that self-repair is the topic, the term "repair" suffices. Levelt (1983) distinguishes between *covert* and *overt* repairs. Covert repairs are characterised by simple interruptions with an editing term (usually (72%) *"uh"*, with covert repairs in Levelt's corpus) and no alteration to the speech already produced, or repetitions of one or more words: it is assumed that such phenomena are the result of monitoring "before the utterance is overtly expressed" (p.55). Overt repairs are changes to speech already produced, either to alter the message to correct an error or to modify the message or completely change tack.

The term "repair" represents a speaker-centred view of the phenomena. The aim of this thesis is to look at the phenomena which result from repair from the point of view of the *listener*. For the listener, it is not of immediate importance to know the precise function of the repair (for example, whether the speaker is qualifying, substituting or correcting an error in their speech). It is more immediately important for the listener to be able to detect that the speech is no longer proceeding fluently and that the speech after the interruption does not follow coherently from the speech before. Repaired speech, whether overt or covert, *is* disfluent: all the categories of repair described by Levelt (1983) result in a break in the normal flow of speech. But not all breaks in fluency are necessarily caused by repair: factors external to the speech production mechanism but still speaker-internal (hiccoughs, coughs, laughter, etc.) may cause speakers to produce incoherent or faltering speech; speaker external factors (distractions, interruptions by other speakers) may do the same. For these reasons we prefer to use the term "disfluency" to cover all the phenomena which constitute breaks

in fluency.

The choice of "disfluency" as a generic term is not entirely uncontroversial. Some recent work in speech production research has used the term "disfluency" in a different way. In discussing what causes the speaker to produce the type of phenomena covered by Levelt's "covert repair", Postma *et al.* (Postma *et al.*, 1990; Postma *et al.*, 1991; Postma & Kolk, 1992; Postma & Kolk, 1993) use the term *disfluency* to distinguish

> "interruptions of a speech plan rather than deviations from this plan",

examples of which are

> "filled and silent pauses, repetitions of words or longer utterance parts, repetitions of syllables and single phonemes, sound prolongations, and blocks (abrupt halting of the speech)",

from *self-repairs*, which are instances of

> "speakers' backtracking in an utterance to correct a speech error or unintended meaning". (Postma *et al.*, 1990, p. 19)

They conclude that what they call "disfluencies" are likely to be the result of "covert repairing", or prearticulatory editing, of internal speech errors. They also suggest that the features of the speech which result from covert repairing can be accounted for by a combination of repair principles which arise from observations about overt repairs. The particular choice of the term "disfluent" to refer just to a part of the set of breaks in fluency seems infelicitous given our reasoning above, but the authors are partly interested in developing a theory to explain stuttering, so their usage may stem from its common employment in the field of speech pathology.

So, of all the terms in the literature used to describe the phenomena which form the raw materials for this study, "disfluency" is preferred for this study, as the most general and the most appropriate from the point of view of perception of spontaneous speech.

2.1.2 Types of Disfluency

Just as the diversity of approaches to the study of disfluent speech produces many different generic terms for the phenomena, so the division of the data into types results in many divisions and many names. In this section, we survey some of these divisions and types.

In an early description of "Hesitation Phenomena in Spontaneous English Speech", which looks at the distribution of disfluencies in stretches of speech of 80 words or more by participants at a conference and draws some preliminary inferences about the nature of encoding units in speech production, Maclay and Osgood (1959) define four basic types of disfluency.

1. **Repeats:** all repetitions from the length of a phoneme to several words that were "judged to be non-significant semantically" (p.24)

2. **False Starts:** all incomplete or self-interrupted utterances. These are subdivided into **retraced** and **non-retraced** false starts, depending on whether the speaker "backed up in an attempt to correct one of the words he had already used" (p.24).

3. **Filled Pauses:** All occurrences of the English hesitation devices [eh, ae, r, @, m].

4. **Unfilled Pauses:** "silence of unusual length and non-phonemic lengthening of phonemes", decided subjectively by the authors.

These basic categories are accepted by subsequent studies (e.g. Martin and Strange, 1968; Duez, 1993; Deese, 1980, who categorises retraced false starts separately as "corrections").

Blankenship and Kay (1964) study the distribution of seven types of hesitation phenomena, omitting "non-phonemic lengthening of phonemes" (of which they find no instances in their data) and "unfilled pauses". Their explanation for omitting the latter is that their study is only concerned with syntactic matters. Their categories are:

1. **Non-lexical intrusive sounds:** Maclay and Osgood's "filled pauses";

2. **Sentence Correction**, where the speaker changes the syntactic plan without restarting the whole sentence anew;

3. **Word Change**, where the speaker substitutes one word for another of the same lexical class;

4. **Repeat:** one or more repetitions of one or more complete lexical items ;

5. **Stutter:** repetition of a unit smaller than a lexical item one or more times;

6. **Omission** of part of a word (strictly speaking, words left incomplete, rather than words with random bits missing!);

7. **Sentence Incompletion:** Maclay and Osgood's False start with no retrace – the speaker breaks off mid-sentence and begins a different sentence.

In studies based in the field of psychotherapy, Mahl (1957; Kasl and Mahl, 1965) distinguishes eight categories of "disturbances and hesitations": *"ah"*, sentence change, repetition, stutter, omission, sentence incompletion, tongue slip, "incoherent sound". The same set of categories is used in later work by Scherer in describing "speech discontinuities" (Scherer, 1979).

In a sociolinguistic approach to conversation analysis, Allen and Guy identify four kinds of "fragmentation" and three main functions. The types they identify are defined as follows:

1. **Incomplete thought**, where the "thought" presented is "too incomplete to be understandable":

   "I just / I think you miss a lot on campus life when you commute".[1]

   The authors seem to mean that the speaker decides to begin a completely different sentence in these cases, but it is not entirely clear from their examples. Their first example also meets the following definition:

2. **Incomplete Word or Phrase**, an example of which is

   "With just a / Just a BA degree",

[1] All examples and quotations are from Allen and Guy, 1974, pp.170-171.

where "the phrase *'Just a BA degree'* does convey a thought, while the partial phrase *'With just a'* calls for something more"

3. **Repetition of words or groups of words which are incorporated in assertions:** once again, the example leaves the reader confused:

   "That's / that's / I'm doing an attitude study on that.

4. **"ah", "eh", "er", "aw" and "uh".**

The functions are described as "filler function", "channel management" and "verbal catharsis". The **filler function** is the use of the use of the fillers ("ah", "eh" etc.) to maintain an even flow of vocalisation while the speaker tries to select a word or form a phrase. **Channel management** involves the use of fillers either to signal the speaker's desire to retain the channel (i.e. hold the conversational turn) or to transfer it to the other participant in the conversation. **Verbal catharsis** is the repairing or restoring of erroneous, misleading or undesired (by the speaker) portions of speech.

Schegloff *et al.* (1977) and Schegloff (1979) do not provide a detailed taxonomy in their studies, but point out a type of disfluency which is often missed in other studies. **"Transition space repairs"** are repairs which take place after possible completion of a conversational turn:

> *"I mean y'know they put up y'know that kinda paper 'r stuff .. the brown paper"* (Schegloff *et al.* (1977), p366 §, 3.11.)

Hieke (1981) presents a "content-processing view of hesitation phenomena". On the basis of the notion of quality control in speech production, he posits two major categories of disfluency: **stall** and **repair**. The categories correspond to speakers forestalling or committing (and subsequently repairing) errors. Hieke points out that the "traditional" categories of disfluency, repeats and false starts are not mutually exclusive and that they have various features in common: they both interrupt the sequence of speech output planning and both require the speaker to backtrack. He observes that repeats may fall into two categories: **Prospective repeats** have a similar rôle to silent and filled pauses and lengthening ("prolongations") – they allow time for lexical search; **Retrospective repeats** are seen as having a bridging function, connecting the restarted speech

to the prior speech which has become separated after a break, reestablishing fluency. Hieke cites Dickerson (1971), who sees this type of repetition as a being a consequence of pausing, where the speaker finds it necessary to restart the current syntactic constituent where a pause has become too long for cohesion to be maintained . The same phenomenon is often heard where a speaker is interrupted by another conversational participant and tries to maintain their turn. One of Hieke's examples (with pause-length in milliseconds in parentheses) shows how the speaker restarts after a pause of about 3 seconds rather than continuing the sentence directly:

"der Vater [640] hm [2240] der Vater [240] ist mürrisch ..."

Hieke's taxonomy thus has the following structure:

1. **Stalls**, which consist of: *silent pauses, filled pauses, prospective repeats, syllabic prolongations*;

2. **Repairs**, which consist of: *false starts, retrospective repeats (bridging).* Repairs are further subdivided to reflect the type of repair action involved:

   (a) **Phonology** repairs are corrections of pronunciation errors;

   (b) **Syntax** repairs are corrections of syntacto-semantic errors (substitution, addition, restructuring); lexical substitutions appear to be included in this subcategory;

   (c) **Rhetoric** repairs are corrections to cohesion (bridging).

In an influential article on speech production, Levelt (Levelt, 1983) presents a detailed taxonomy for disfluencies (self-repairs), designed to account for speakers' motives in making repairs. On the basis of speaker-motive, he distinguishes 5 major categories of repair.

1. **D-Repairs:** these occur where the speaker changes their mind about the current message and decides to say something **D**ifferent.

   "We go straight on or ... We come in via red, then go straight on to green"[2]

[2]Examples based on Levelt's work are adapted from the English gloss of the original Dutch.

2. **A-Repairs:** here, the speaker realises that the information in the intended message needs to be qualified in some way in order to make it more **A**ppropriate to the context. Four types of A-repairs are identified:

   - **AA-Repairs:** these are concerned with undoing **A**mbiguity of reference.

     "We start in the middle with ... in the middle of the paper with a blue disc"

   - **AL-Repairs:** these adjust the **L**evel of precision needed for accurate description of a concept (usually moving from a less to a more precise term).

     "... with a blue spot ... a blue disc at the upper end"

   - **AC-Repairs:** these adjust for **C**oherence with the previous text.

     "... you go one up, there's uh ... you come to yellow"

   - **ALC-Repairs:** in some cases it was impossible to determine whether the speaker was repairing for level or for coherence: such cases were coded ambiguously.

3. **E-Repair:** the speaker discovers that what they have said contains an **E**rror, rather than being merely inappropriate. Error repairs fall into three subcategories:

   - **EL-Repairs:** the speaker changes some lexical item (of any lexical class).

     "... straight on red, or sorry, straight on black"

   - **ES-Repairs:** the speaker begins a **S**yntactic construction but can not end it properly, so begins a new one to replace it.

     "... and black to ... from black to ... right to red"

   - **EF-Repairs:** **Ph**onetic repairs, rather infrequent, despite the volume supplied by the speech error literature.

     "... a unut ... unit from the yellow dot"

4. **C-Repairs:** where the speaker just interrupts their utterance with an editing term (filled pauses as well as *"rather"*, *"well"* etc.) or repeats one or more lexical items without changing, adding or deleting anything. Levelt assumes that covert repairs are evidence of the ability of the speaker to attend to "inner speech" of some kind and to make changes to the speech before it is articulated.

5. **R-Repairs:** this category is for the **R**est of the repairs, which cannot be fitted into any of the previous 4 categories (only 2.5%) of Levelt's data falls into this category.

As already mentioned (section 2.1.1), Postma *et al.* (Postma *et al.*, 1990; Postma *et al.*, 1991; Postma & Kolk, 1992; Postma & Kolk, 1993) use the term "disfluency" to refer to Levelt's covert repairs, while pursuing the goal of discovering if such "repairs" really are "covert". Another study which makes use of a modified version of Levelt's classification is by Blackmer and Mitton (1991) . These authors make five alterations to Levelt's scheme.

1. A-repairs and D-repairs are maintained as separate categories, but subsumed under a new category of **conceptually-based repairs**. This type of repair is assumed to correct errors that originate in the component of the speech production mechanism referred to as the "conceptualiser".

2. E-repairs are renamed **production-based repairs** to reflect the theory that the problems they deal with originate in the formulator or articulator.

3. Levelt's subcategories of A-repairs and E-repairs are removed and new subcategories of the A-repairs created. These new subcategories represent A-repairs which replace prior speech – **appropriateness replacements** – and those which insert new speech – **appropriateness inserts**, an idea based on Schiffrin's (1987) distinction between replacement and background repairs.

4. C-repairs are given three subcategories, reflecting their three possible realisations, entertaining the possibility that the forms might reflect different

processes: **covert-repairs with an editing term**; **covert repairs with repetition**; **covert repairs with an editing term and repetition.**

5. A further subcategory for C-repairs allows categorisation of filled pauses between utterances.

Levelt's R-repair category is retained but restricted to use with overt repairs.

Van Wijk and Kempen (1987) , although examining Levelt's theories, choose to distinguish between **retracing** and **non-retracing** repairs. This dichotomy is made rather than a reference to error and appropriateness repairs, since their interest is more in the formal relationship of the repair to the speech before the interruption than to speakers' motives for making the repair. They also distinguish two mechanisms for performing repairs: **reformulation** and **lemma substitution**. In reformulation, the important linguistic unit is the major syntactic constituent. In lemma substitution, a prosodic unit, the phonological phrase is of more importance.

Allwood *et al.* (1987) approach the topic from the point of view of pragmatics, under the heading of "Speech Management Phenomena" (SM), and give a very thorough account of the structure and function of the phenomena. They distinguish Basic SM Expressions and Basic SM Operations as basic SM features.

1. **Basic SM Expressions**

   (a) pause (marked "//" and signifying lack of speech and gesture while holding a turn);

   (b) simple SM expressions (filled pauses, *"eh"*, *"äh"*, *"m"*)[3];

   (c) explicit SM phrases (e.g. *"vad heter det" (what's it called)*);

   (d) other SM sounds (e.g. smacking, sighing, hissing and other sounds which are difficult to classify);

2. **Basic SM Operations**

   (a) lengthening of continuants;

[3]Swedish data.

(b) self-interruption;

(c) self-repetition;

These features can occur alone or in combinations. In addition, the authors describe **Complex SM Operations**:

1. **Holistic Operations**, where an interruption is followed by a resumption involving one of the following operations:

   (a) Deletion – the resumption repeats part of the original message, with a fragment, word or phrase removed;

   (b) Insertion – the resumption repeats part of the original message, with a word or phrase added;

   (c) Substitution – the resumption repeats part of the original message, with a word or phrase changed;

   (d) Reordering – all the constituents of the original message are repeated, but in a different order.

2. **Integrated Operations**, where a Basic SM Expression marks a function (for example lexical search) which can be left unmarked by a holistic operation. In their example (20) (Allwood *et al.*, page 22):

   för att inte **ääh // eh** *för att hålla en del gröder vid liv*

   (*in order not to* *in order to keep some crops alive*)

   Basic SM Expressions (**ääh // eh**) are *integrated* as a suboperation within the holistic operation of substitution.

3. **Linked Operations** occur where complex operations occur in sequence with other SM features but not as integrated operations. Three subcategories are defined:

   (a) **Recursive Linking**, where one SM feature is embedded within a holistic operation but is not a suboperation of it:

   om de– **om: varje // chartist** *får ett eh större antal anhängare ...*

   (*if it* *if every* *chartist gets a greater number-of adherents*)

(b) **Conjunctive Linking**, where an SM feature is sequentially linked to a holistic operation. The example assumes that the insertion operation is initiated before the pause, rather than after it:

en liten risk // **väldi liten risk**

(a small risk very small risk)

(c) **Overlapping**, where two or more holistic operations occur in the same stretch of speech without operating on precisely the same structure:

nä men de drabbar ju **de kan drabba** *sådana områden där ...*

(no but it strikes you-know it can strike such areas where)

here, the deletion of **ju** overlaps with the substitution of **kan drabba** for **drabbar**.

In comparison to other accounts, Allwood *et al.* provide a more complete account of the phenomena and do so from a more neutral perspective than, say, Levelt's speech-production based account.

Approaches to disfluent speech in computational linguistics which concern themselves with types of disfluency are concerned with finding formal aspects of the speech which can be used to identify and remove the extra material in order to achieve a correct parse.

Hindle (1983) , while not explicitly defining the types which his algorithm handles, refers to two general types. His algorithm looks for **repeated matching surface strings** (repetitions) and **repeated matching syntactic categories**. To distinguish disfluent repetitions of strings and structures from fluent and syntactically acceptable strings, Hindle introduces the notion that disfluencies are *always* marked by a discrete **editing signal** at the interruption point. This feature of disfluent speech, suggested by Labov (1966), is described as a *"markedly abrupt cut-off of the speech signal"*. Amongst other authors surveyed, only Taylor and Cameron (1987) assume the presence of such a powerful signal, although Schegloff (Schegloff *et al.*, 1977; Schegloff, 1979) lists a **"cut-off"** (a glottal or other stop) as one of a number of potential initiators of repairs. Hindle places part-words (fragments) in the same **non-lexical** category as filled pauses, which are filtered out before the other categories are identified (so that a fragment can

not be seen as, for example, a repetition in this analysis). A separate treatment is envisaged for another category, the **restart**, which Hindle sees as

> "less sensitive to syntactic structure and flagged not only by the editing signal but also by a lexical item" (Hindle, 1983, p.127).

A restart is the beginning of a new sentence. The lexical item is one of the set which includes *"well, ok, see, you know, like I said,"* etc.[4].

A more recent study in the field of computational linguistics (Bear *et al.*, 1992; Shriberg *et al.*, 1992) has defined types of disfluency more formally. These authors define the disfluencies in their study in two ways, first, in terms of the **length** in words of the string which is to be "deleted" for a correct parse to be achieved and second, as one of 5 basic types: **fragments, repeats, insertions, replacements** and **other**.

2.1.3 The structure of Disfluency

It is helpful in discussing disfluency to be able to refer to various features in the structure of the utterance in which it occurs. With the central point being generally referred to as the **interruption**, it is useful to be able to refer to the speech before and after the interruption, to the spoken material that is replaced and to the speech which replaces it as well as to the previous, intervening and following speech. Not many authors describe the structure of disfluency taking these factors into account to any great degree.

Taylor and Cameron (1987) refer to the speech on either side of the interruption as the "pre-discontinuity" and "post-discontinuity". Hindle (1983) refers to the speech that is replaced when a speaker repairs as the "expunction site", because all the material in that portion of the speech is expunged from any further syntactic analysis.

The most thorough description of the structure of disfluency at this level comes from Levelt (1983). He identifies three major areas in a disfluent utterance: the **original utterance** (OU), the **editing phase** and the **repair** (R) itself.

[4]As will be shown in Chapter 3, the use of such lexical items is much less common than Hindle hopes.

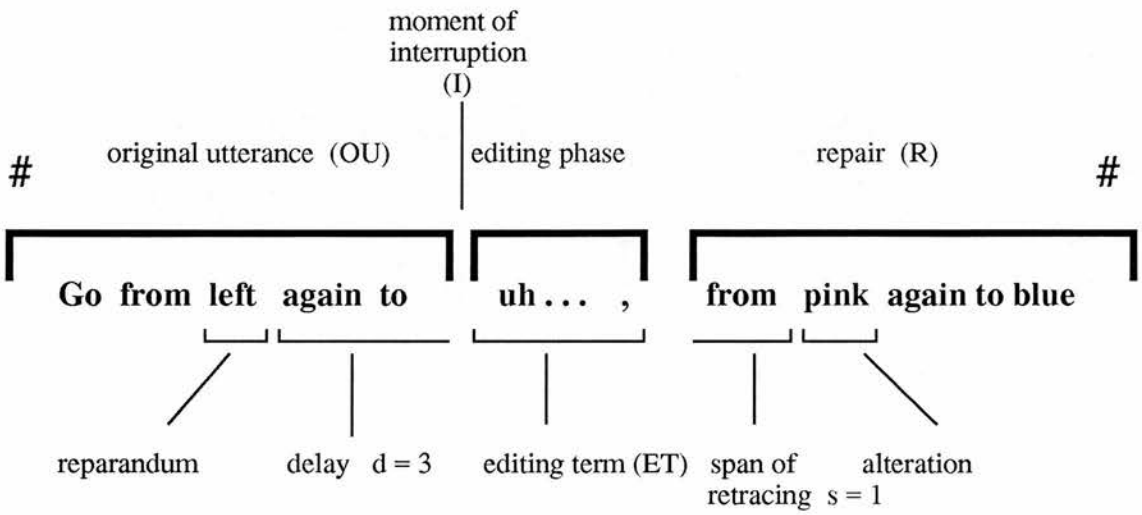

Figure 2.1. Levelt's structure of repair. From Levelt (1983), page 45.

The OU consists of all the speech from the sentence onset to the interruption (I) and contains the item that is to be repaired, the **reparandum**. Any speech between the reparandum and the interruption point is referred to as the **delay** of interruption (which is given a value corresponding to the number of syllables it contains). The editing phase is a period of variable length, which may or may not contain an editing term (*"uh, rather, well"* etc.). Within the Repair is the **alteration** (which "replaces" the reparandum). The alteration may be preceded by some **retracing**, which repeats words from the OU prior to the reparandum: the span of retracing is measured in syllables. An example from Levelt's article is shown in figure 2.1

In recent work, Nakatani and Hirschberg also divide the disfluent utterance into three intervals: the **reparandum interval** the **disfluency interval** and the **repair interval**. Their reparandum interval includes all the material that the speech after the interruption effectively replaces (i.e. from *"from"* to *"to"* in the OU in figure 2.1); the disfluency interval includes all silence and pause fillers and cue words from the offset of the reparandum interval to the onset of the repair interval; the repair interval includes all speech from the resumption after the disfluency interval to the end of the material which "replaces" the reparandum (i.e. from *"from"* to *"to"* in the repair in figure 2.1).

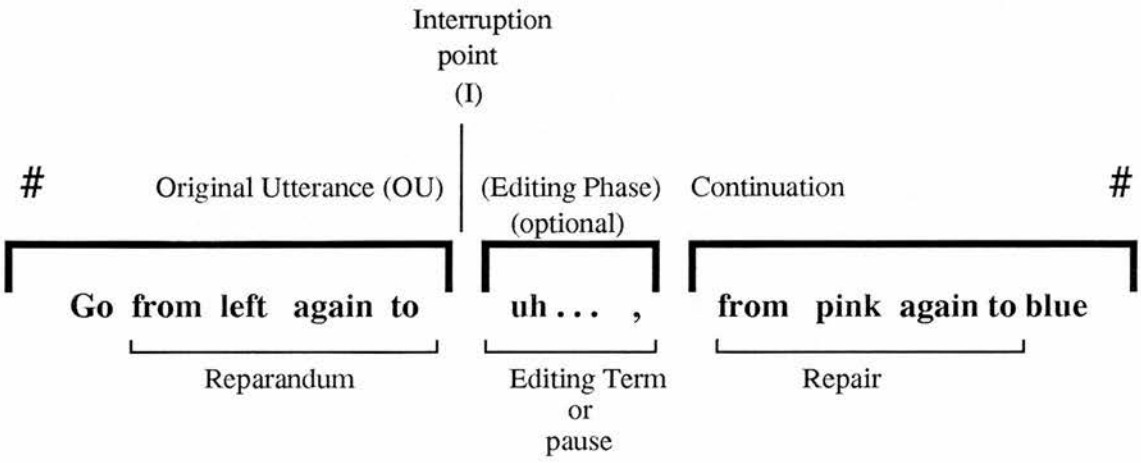

Figure 2.2. The structure of disfluency. (Adapted from Levelt (1983), page 45.)

2.1.4 Terms and Types in this Thesis

Having surveyed the possible choice of terminology and typology used in describing disfluency, this section is concluded with an explanation of terminology we will use to describe the data in the present study.

The main criteria in deciding on terminology and typology are that they should be kept fairly simple but cover all of the data in an intuitively understandable way. To begin with, the structure of disfluency is described following Levelt (1983), but in a slightly simplified form (compare figures 2.1 and 2.2). The three main phases in a disfluent utterance are now called the Original Utterance (OU), the Editing Phase and the Continuation. The OU contains the reparandum, as in Levelt's description, but the reparandum consists of all the speech that is effectively replaced by the repair, including Levelt's "delay". The Continuation begins at the onset of the Repair and runs to the end of the utterance. The Repair is just the speech which Levelt's diagram labels as "span of retracing" and the "alteration". The Editing Phase is optional: in its absence, the interruption point is the onset of the repair.

Few criteria based on the perceptual properties of disfluency can be suggested, since so little is known about the perception of different types of disfluency, but the typology should take account of factors like the general structure of the disfluency, which may affect the perception: word repetition might hold different

prosodic information from word substitution; longer reparanda might result in larger differences in pitch at interruptions; reparanda ending in fragments might provide different cues from reparanda ending in full words.

A fundamental distinction is made between disfluency which makes no difference to the syntactic coherence of the utterance and that which adds words or fragments which would have to be removed for a successful parse to be found. The former case consists of disfluencies which are just **silent pauses** or **syllabic lengthening** (or "stretch" (Schegloff *et al.*, 1977; Schegloff, 1979)) or **filled pauses** at locations where the syntactic and prosodic context does not predict them (equivalent to Maclay and Osgood's (1958) categories of unfilled and filled pauses). The latter include the two major types, **Repetition** and **False Starts**[5]. Repetition includes all strings of words or fragments, or words and fragments which are repeated verbatim and without major alterations to stress. Subcategories of repetitions identify the number of times a fragment, word or phrase is repeated in one episode. Levelt's definition of "covert" repair includes hesitation repetitions; Hieke (1981) shows that there are two orientations for repetitions – prospective and retrospective (Page 12): in this analysis, all repetitions are categorised together, and all disfluencies which involve any repetition or retracing will be referred to as "repairs". False starts will be subcategorised according to the type of alteration that is made to the OU. Four types of alteration are identified:

1. **Word Change:** a lexical item in the OU is substituted;

   drinking so quickly in those *last / that last half hour*

2. **Qualification Change:** the speaker modifies something they have said by adding or removing a qualifying word or phrase:

   it doesn't */ I think it doesn't realise*

3. **Pronunciation Change:** a word is altered because of an articulation error or inappropriate stress:

   many people in the west have grave reve- */ reservations*

[5] Although, formally speaking, repetitions are a subset of false starts, since both types involve a restart, it is convenient to separate the two categories.

they would / *they'd do introductory Latin and Greek*

4. **Syntactic Change:** this includes mid-sentence alterations as well as complete restarts:

   so anyway I uh / *are you going tomorrow*

In addition to these major categories, all repairs are subcategorised for the length of the reparandum. Three categories based on the number of whole words are identified: reparanda of one whole word, of two whole words and of three or more whole words. Three more categories differentiate between reparanda consisting of just a single fragment (which may consist of just one phonetic segment or of a syllable or more), one whole word plus a fragment and two or more whole words plus a fragment.

2.2 Approaches to Processing Speech with Disfluency

Until recently, most work in speech understanding has made use of "lab-speech" – careful recordings of specially constructed words or sentences – which contains none of the disfluency so common in spontaneous speech. As a result, little is known about how the human processor understands disfluent speech and only a few studies have approached the problem from the point of view of automatic speech recognition. In this section, we describe some of the most important work on disfluent speech in the fields of first psycholinguistics and then computational linguistics.

2.2.1 Psycholinguistic Approaches

The specific research area of this thesis, the detection of disfluency in spontaneous speech by human listeners, has been largely neglected. As a result, there are very few published studies of direct interest and few foundations on which to start to build a psycholinguistic model. Most of the psychological interest in disfluency has come from the field of speech production. For example, studies of speech

monitoring for self-correction of errors have looked at how features of disfluent speech reflect the production processes involved in repair (Levelt, 1983; Levelt, 1989; Blackmer & Mitton, 1991; Postma *et al.*, 1990; Postma *et al.*, 1991; Postma & Kolk, 1992; Postma & Kolk, 1993), although Levelt does spare a thought for the listener in his article (1983). The studies discussed are few and there is still a great deal of work to be done in human processing of disfluencies.

In one of the earliest studies on the perception of disfluencies, Martin and Strange (1968) found that listeners could not accurately reproduce disfluencies orally or mark them on transcripts. They found a tendency to displace the disfluencies they heard to clause boundaries. In different experimental conditions, subjects were given a range of different tasks, from simply being asked to repeat what they heard ("ordinary encoders") to being asked to repeat everything verbatim ("exact encoders"), and being trained to understand what constituted disfluency in the signal. The rate of accurate identification ranged from 6% for ordinary encoders to 15% for exact encoders. The instruction to attend to disfluencies had the effect of improving their detection but at the cost of reducing the percentage of words in the main message that were reproduced correctly. The difficulty that transcribers have in detecting, placing and reproducing disfluencies accurately is well-attested: for example, researchers on speech disorders report that the questionable trustworthiness of their data is one of the fundamental problems for their field (Moore & Perkins, 1990; Perkins, 1990; Aram *et al.*, 1991; Cordes *et al.*, 1991; Ingham & Cordes, 1992; Ingham *et al.*, 1993).

Howell and Young (1991) suggest that some features of disfluencies help listeners to detect the error: their study used synthesised speech and varied pause at interruption point as well as "added stress" on the repair. Their materials were all artificially created repetitions and alterations. Their experiments used two techniques: a comprehensibility judgement task, where listeners were asked to judge which of two sentences would be easier to comprehend if it were produced in real speech; a reaction time test on a reproduction and editing task, where subjects were asked to repeat disfluent sentences as soon as possible after they had heard the prompt, without reproducing the disfluency in the sentence. The comprehensibility tests and the reaction times both supported the hypothesis that pause and "added stress" are useful for listeners in indicating the presence of

disfluency. It is not clear that these features aid detection in their complex tasks rather than some later process of comprehension or introspection. The results might be as easy to attribute to the effects of chunking on rehearsal, for example, as to the process of detecting and correcting a disfluency. Moreover, the premise that "added stress" and pause are always present in disfluency, or even present in the majority of cases, is not even supported by their own data, so that the generalisability of their results to all disfluency is in doubt.

Duez (1993) finds that "prepausal" vowel lengthening is a crucial factor in the detection of interruptions in spontaneous French speech, where interruptions include disfluencies of all kinds, as well as fluent clause boundaries. In experiments with speech from a corpus of political speeches, political interviews and casual interviews with politicians, subjects were asked to press a button each time they heard an interruption in the speech. The tapes were presented in two conditions: normal and with spectrally inverted-speech. Spectral inversion (Blesser, 1972) in the frequency band 200-4000Hz results in a signal which makes segmental information impossible to perceive but maintains suprasegmental information in the form of the fundamental frequency and relative amplitude information. "Pauses" (disfluencies) were counted as "Subjective pauses" where they were reported at least twice (i.e. in both presentation modes by one subject, or by two subjects in any mode) and where the "pause" did not occur at a syntactic boundary immediately preceded or followed by a stop consonant. The most important acoustic factor in these perceived pauses is found to be the presence of prepausal lengthening. Of the 32 disfluencies ("hesitation pauses") detected by Duez's subjects, 11 had a lengthened syllable, 8 a filled pause and 11 had both lengthening and filled pause. But no significant effect of disfluency on detection of pauses was found – the presence of disfluency *per se* did not correlate with instances of pause judgements.

2.2.2 Computational Linguistics

There is a thin but fuzzy line between computational linguistics and natural language understanding in artificial intelligence. In this survey I will use the general heading "computational linguistics" (CL).

Carbonell and Hayes (1983) suggest recovery strategies for parsing input which contains disfluencies: like most accounts in CL, they make the unsound assumption that the problem of word segmentation and recognition has been handled and the problem is reduced to that of deleting inappropriate words or syntactic constituents (the problem is discussed under the heading "Spurious constituents"). They suggest three methods for for detecting the "spurious" part of "broken-off and restarted utterances":

1. where a sequence of two constituents of identical syntactic and semantic type is found where only one is permitted, ignore the first one;

2. recognise explicit corrective phrases (such as *"I mean"*) and if the constituent to the right is of the same syntactic and semantic type as the one to the left, substitute the right constituent for the left one;

3. in making such substitutions select the minimal constituent on the left to be substituted. In:

   > *"Add a high speed tape drive, that's disk drive, to the order"* (Carbonell and Hayes, 1983, p.128)

   "disk drive" should substitute *"tape drive"* and not *"high speed tape drive"*, which has the same semantic and syntactic type.

The article presents a variety of other methods for handling a wide range of types of "extragrammatical language", including typing and spelling errors, but does not address the issue of what non-syntactic cues may be available, since the parser is assumed to have syntactically and semantically labelled words as input.

Langer (1990) also groups disfluent speech with other ill-formed utterances when he proposes some parsing techniques to handle **"explicit repair"** and **"ungrammatical repetitions"**. Explicit repair is a repair which contains a "repair indicator" at the interruption (examples given are *"uh no"*, *"nonsense"*, *"sorry"*). Where a repair indicator is found, the strategy is to peel words off the end of the speech prior to the indicator one by one, trying to parse the resulting sentence each time, until an acceptable parse is found. As a result of this strategy, an utterance like

"you put the left one uh the red one to the left"

has to be checked three times, as the words *"one"*, *"left"* and *"the"* are removed one by one, before the correct utterance is revealed. But a similar utterance might cause this system to produce incorrect output. The utterance

"you put the left one uh red one to the left"

would result in the semantically anomalous

"you put the left red one to the left".

Ungrammatical repetitions include simple repetitions of identical strings, which are dealt with by simple removal of the first instance, and incomplete repetitions and repetitions which introduce new lexical items. To detect these, the input string (all words assumed recognised and labelled) is scanned for two different occurrences of the same lexical item (which may differ in inflectional properties); if they are found, the substring starting with the first instance and ending with the word before the second instance is parsed, and, if possible, assigned a syntactic category symbol; the following string (commencing with the second instance of the repeated word) is then parsed; if this parse results in the discovery of a constituent of the same syntactic category as was found for the string containing the first instance of the word, then the string containing second instance is considered to be a suitable replacement, providing it forms a complete sentence with the rest of the utterance. As an example, in the input string

"some blocks some red blocks are small",

the word "some" is the first to detected as a repetition; the substring "some blocks" is an NP; the string "some red blocks" is also marked as NP; the sentence "some red blocks are small" is found to be grammatical, so the parse succeeds.

A different CL approach to processing spontaneous speech and handling disfluencies and other "noise" phenomena is to avoid them altogether. Luzzati (1987) describes a skimming parser for use in a strictly limited environment (train timetable enquiries). Disfluencies are part of the "syntactic noise" category which this style of parser effectively ignores by skimming the input for keywords in known contexts.

In contrast, Hindle's paper (1983) breaks the one of the moulds of CL approaches in that it treats disfluency as a separate set of phenomena from other types of grammatical deviance. He describes an editing system which works in tandem with Marcus' deterministic parser, *"Fidditch"* (based on processing principles in Marcus, 1980). The system takes as input strings of transcribed words and depends crucially on the presence of an *"editing signal"*, at the interruption point in disfluent utterances (see page 18). This signal, assumed to be a *"phonetically identifiable signal placed at the right edge of the potential expunction site"* (Hindle, 1983, p.128), triggers a set of three copy editing rules, which find two elements on either side of the editing signal which are copies at some level of description and expunges, or deletes the first of the two instances. These copy editors operate as follows:

1. **Surface Copy Editor:** a non-syntactic rule which matches surface strings on either side of an editing signal and removes the first instance. In Hindle's system, this applies to surface strings – orthographic transcriptions of words – before any syntactic processing has applied. So in this example, the first instance of *"if they'd"* and the portion *"I wou-"* are expunged before parsing begins:

   > *"Well* if they'd - - *if they'd had a knife* I wou - - *I wouldn't be here today"*[6]

2. **Category Copy Editor:** if two matching syntactic constituents in the parser's buffer of complete constituents are found on either side of an editing signal, the first is expunged. Thus, in the following example, *"that"*, marked as a determiner, is expunged in favour of *"the"*, and the first of two verbs, *"have"* is also expunged:

   > *"I was just* that - - *the kind of guy that didn't* have - - *like to have people worrying"*

3. **Stack Copy Editor:** if the first complete constituent in the parser's window is preceded by an editing signal, the Stack Copy Editor looks for an

[6]Examples are from Hindle, 1983, p.125. The **editing signal** is marked "- -".

incomplete constituent with the same label in the parser's push-down stack of incomplete constituents. Any copy found there is expunged, along with all descendants of that constituent. For the example

"I think that you get - - *it's more strict in Catholic schools"*,

the incomplete embedded sentence *"you get"* is expunged.

Hindle comments that the Surface Copy Editor may be functionally redundant, given that the syntax-based editors would usually be able to expunge any surface copies using syntactic criteria. But he attaches psychological importance to the editor and the algorithm in general by stating:

"...it seems that the Surface Copy Editor must exist at some stage in the process of syntactic acquisition. The overlap between it and the other rules may be essential to learning." (Hindle, 1983, p.125)

Sentence restarts are set to one side in Hindle's analysis, because they are seen as being "less sensitive to syntactic structure" (p.127). It is assumed that they are signalled by a lexical item (*"well, ok, you know"*, etc.) as well as the editing signal, and that "specific intonational signals" (which are not, however, specified) are also present. Expunged material is acknowledged as potentially having semantic content, and removed completely only from the syntactic parse.

A test of the system on a transcription of the spontaneous speech of one speaker gives a success rate of 97% in editing out disfluencies.

Hindle thus presents an editing system consisting of three editors which remove various types of disfluencies from a transcription of spontaneous speech. The crucial element in the system, as we have seen, is the editing signal. Although an attempt to define it is made in the article, no serious definition has been found, and no researchers since Hindle have been able to define such a signal. Should these failures indicate that there is no such signal, Hindle's system offers no more than Langer's. The first experiment described in this thesis represents an attempt to test the psychological reality of the notion (Chapter 4).

More recent CL approaches to disfluent speech have responded to this hiatus by eschewing (Bear *et al.*, 1992; Shriberg *et al.*, 1992) or adapting (Nakatani & Hirschberg, 1993a; Nakatani & Hirschberg, 1993b) the notion of editing signal.

The SRI work (Bear *et al.*, 1992; Shriberg *et al.*, 1992) uses pattern matching followed by syntactic, semantic and acoustic analysis to detect and correct disfluencies.

1. The **Pattern Matching** component marks two types of event:

   (a) identical sequences of words;

   (b) simple syntactical anomalies, like illegal pairs of nonidentical determiners or prepositions.

   The output of the Pattern Matching component is the input to the linguistic analyses, which attempt to distinguish true disfluencies from *false positives* – repeated strings which are intended and fluent – (*"flights for <one> one person"* vs. *"US Air flight one one five"*).

2. The **Syntax and Semantics** of the Pattern Matcher's output is examined: if a parse succeeds, the sentence is marked as a false positive; if the parse fails, pattern matching techniques detect repairs by looking for any of a set of patterns which represent repeats, insertions or substitutions (mentioned on page 19, above), and the appropriate edit is performed before a second parse is attempted.

3. **Acoustics:** certain acoustic features of the potential repairs are examined to establish their use in distinguishing disfluent sentences from false positives:

   (a) Duration, pause duration and F_0 values are found to be of use in distinguishing certain patterns of disfluency from their false positive pairs;

   (b) cue words (*"well"* and *"no"*) in repairs were found to differ from the same words in fluent speech in terms of the direction of F_0-movement, presence of lexical stress and continuity with the surrounding speech (see section 2.3.2);

   (c) fragments were found to confuse the word recogniser in two ways: they could either be recognised as full words on their own or be recognised

as part of a neighbouring word. Glottalisation is seen as a possible acoustic cue to the presence of vowel-final fragments.

In tests of the system, the pattern-matching component correctly identified 76% of disfluencies and incorrectly hypothesised a number more, giving an overall precision of 62%. The syntactic and semantic analysis of the output of the pattern matcher resulted in 57% of the disfluencies in that output being correctly marked.

In another study which takes acoustic information into account, Nakatani and Hirschberg (1993a, 1993b) compare various acoustic and prosodic features of repairs with similar features at fluent phrase boundaries. They attempt to distinguish word pairs on either side of interruption sites from word pairs across phrase boundaries by taking into account a combination of 16 features, including the presence of filled pauses and fragments, F_0 and amplitude values, pause durations, the presence of stress, as well as some simple lexical pattern matching strategies. They report a success rate of 78-83% in distinguishing disfluent interruptions from intonational phrase boundaries with 89-93% precision on a subset of their data, 202 disfluent utterances containing 223 repairs. The higher success rate is found when the feature "fragment" is included in the analysis (74% of their repairs have reparanda ending in fragments (see page 37)). When "fragment" is omitted, the most important features in identifying repairs are pause duration, lexical matching and the distance in words from the beginning and end of the utterance.

Discussion

While some workers in CL would like to make claims about the psychological reality of their approaches (e.g. Hindle, 1983), most approaches make the fundamental but unrealistic assumption that the input to the parser is a sentence-length string of isolated words waiting to be integrated into a syntactic structure. The problems of finding words in continuous speech are assumed to be solvable on the basis of the acoustic signal, without top-down information. Work in psycholinguistics would suggest that such an assumption is somewhat naïve in its optimism (Mehler *et al.*, 1981; Bard *et al.*, 1988; Cutler & Norris, 1988; Sebastian, 1992; Quené, 1992; Cutler & Butterfield, 1992; Cutler *et al.*, 1992; Cutler & Mehler,

1993). As a consequence of the approach, however, the problem of processing disfluent speech is seen as essentially a syntactic problem: syntactic patterns are identified which allow disfluent portions of the signal to be identified and the appropriate string to be removed from the parse. Recent studies which take into account acoustic and prosodic properties of disfluency attempt to distinguish disfluent events from specific types of events in fluent speech (SRI's "false positives", Nakatani and Hirschberg's fluent phrase boundaries) rather than attempting an on-line approach.

2.3 Acoustic and Prosodic Cues to Disfluency

Hardly any work in psycholinguistics has looked in detail at the use of acoustic and prosodic cues in the processing of disfluent speech. Approaches by Computational Linguists to processing disfluent speech have concentrated on syntactic cues for the detection and of disfluency and the resolution of associated parsing problems. Acoustic and prosodic information have been largely overlooked until very recently, because the input to parsers is usually assumed to be strings of words from transcriptions. In this section, the possible use of such information in detecting disfluency is examined. Since so little previous work has looked at human perception of disfluent speech, it is appropriate to refer first to some work which looks at the contributions of prosody to the perception of *fluent* speech, before looking at some specific cues suggested by researchers interested in disfluency.

2.3.1 Prosody in the perception of fluent speech

A good overview of some contributions of prosodic factors to the perception of fluent speech was provided by Nooteboom *et al.* in 1978. For these researchers, prosodic continuity had the utmost importance for speech perception:

> "A listener's ability to hear sequences of speech-like auditory events as either having or not having prosodic continuity is probably essential to his ability to perceive speech at all ..." (Nooteboom *et al.*, 1978, p.100).

Two of the main reasons for this are that prosodic continuity is helpful in perceiving different speech sounds as integral parts of meaningful patterns (spectral continuity, for example, allowing listeners to hear sequences of vowel sounds or CV syllables as single auditory patterns (Dorman *et al.*, 1975; Cole & Scott, 1973)) and that it can help the listener attend to a single speech source when several are present (Darwin, 1975).

Darwin's study is interesting not only for the conclusion that it shows that listeners use prosodic information in attending to one speaker in a crowd, but also for the observation that prosodic information overrode semantic information temporarily in a speech-shadowing task. His experiments presented listeners with two different speech signals, stories read by the same speaker, simultaneously, one in each ear. For each pair of passages four recordings were made: two recordings were of the original passages and the other two were made by the speaker smoothly combining the first part of one passage with the second part of the other. The recordings enabled four dichotic listening conditions to be tested, where the listener was instructed to shadow the speech heard in one ear only.

1. *Normal:* the two original passages were paired;

2. *Semantic Change:* the original passages with the smooth combination of the start of one passage with the end of the other;

3. *Intonation Change:* the semantic change passages switched ears after the first half of the passages, so that the passages in either ear were semantically continuous at the break point[7], but intonationally discontinuous;

4. *Semantic and Intonation Change:* the two original passages were switched from one ear to the other at the break point, making the signal in both ears both semantically and intonationally discontinuous.

Two types of error were examined: *omission* errors, where shadowers missed two or more words around the crucial point; *intrusion* errors, where, at the crucial point, shadowers reproduced some speech from the channel (earphone) they were

[7]An acoustically smooth break was ensured by making the changeover point at a stop consonant

told to ignore. The results showed that close shadowers produced significantly more intrusion errors in both conditions where intonation switched than when only semantics changed. In other words, subjects showed a significant tendency to follow the intonation pattern of the speech they were hearing, regardless of the semantic content, for brief periods after the break. From the point of view of disfluent speech, the idea that the listener's attention to prosodic continuity is of prime importance in perception is interesting in that if prodosic *dis*continuity can be defined and found in disfluency then it might be that the prosody of the speech around a disfluency can alert the listener to the presence of the disturbance earlier than syntactic and semantic information.

Darwin also gives evidence that intonational incoherence with the content of a sentence affects the comprehension of a sentence. Using a method similar to that of Wingfield and Klein (1971), he took pairs of sentences with common strings of words and cross-spliced those strings so that two out of four sentences presented in an experiment had abnormal intonation. The recall rate for the abnormal sentences was significantly lower than for the normal sentences. Abnormal intonation clearly had an effect on the processing of the sentences.

Wingfield (1975) showed that prosody aided the comprehension of speech in difficult listening conditions. In his experiments, as the perceived speech rate was raised (via a time-compression method, which increases the speed of the speech without altering the relative timing and prosodic pattern), the comprehensibility of sentences with anomalous prosody decreased significantly faster than that of sentences with normal prosody.

Other work has shown that prosodic information is of use in resolving or avoiding syntactic ambiguity in fluent speech. Duration and pitch can be used to differentiate between two possible structures in a syntactically ambiguous sentence (Lehiste *et al.*, 1976; Streeter, 1982; Scott, 1982). A more recent study (Price *et al.*, 1991) suggests that different types of syntactic structures differ in the degree to which they can be disambiguated by prosodic information.

Studies involving online processing suggest that prosodic cues may be able to resolve local, temporary syntactic ambiguities early in the sentence. Beach (1991) shows that listeners can judge the syntactic structure of a sentence using the prosody at an early point in a sentence that is temporarily structurally

ambiguous. Subjects presented with a sentence onset like

Jay believed ...

(short version) and

Jay believed the gossip ...

(long version), which is structurally ambiguous until near the end of the sentence, were able to judge equally well whether the onset came from a sentence completed by a direct object (*... the gossip about the neighbours right away.*) or a sentence complement, (*... the gossip about the neighbours wasn't true.*), given either the long onset or just the short onset. In a second experiment with synthesised speech, both pitch and duration interacted in influencing the online choice of expected sentence structure at an early point in the sentence. Marslen-Wilson *et al.* (1992) also find that prosodic factors can affect the early stages of parsing and interpretation of attachment ambiguities. Grosjean (1983) also provides evidence that listeners glean information about the rest of the sentence (in this case, how long it will be) from earlier prosody.

Rhythmic information may also be of use in processing speech. Cutler (1976) found reaction times to phoneme targets in a position where a sentence accent was predicted by the preceding prosody were faster than to targets where low stress was expected, even if splicing had made the target words identical. Reaction times to phoneme targets in rhythmically *disrupted* speech (Meltzer *et al.*, 1976; Shields *et al.*, 1974; Martin, 1979) are found to be longer than for the same targets in rhythmically normal speech. Buxton (1983) suggests that listeners are sensitive to rhythmic information to the extent that they can predict the timing of stressed syllables. The results of Meltzer *et al.* and Buxton are disputed by Mens and Povel (1986), who suggest that they may be an artefact of the splicing procedure used which may have introduced phonetic distortion. Their replications of the previous studies produce no evidence for a predictive rôle for rhythm on a sentential basis. Tyler and Warren (1987) find that local disruption of prosody – the insertion of a pause *within a phonological phrase* – increased word monitoring latencies. They also find that global disruptions increased monitoring latencies, which they take to mean that the overall intonational pattern of an

utterance is informative to a listener. Pitt and Samuel (1990) find no evidence that rhythmic expectations build up during the processing of a sentence but do find an "attentional bounce" effect in contexts which had unusually strong rhythm – reaction times to phoneme targets in places where the context predicted that stress would fall were higher than in places where stress was not expected, even if there was actually *no* stress there.

In effect, perceptual studies of prosodic factors in *fluent* speech suggest that listeners pay attention to intonation and stress, making use of such information in several ways. It would be suprising if such information were not also useful for *disfluent* speech. Darwin's finding that prosody can lead speech shadowers off the task of following a particular channel suggests that prosodic information is followed closely, and may be accessed sooner than syntactic and semantic information: we might extrapolate from this that prosodic information could have a key rôle in the detection of disfluency – listeners may be able to detect a breakdown in prosodic continuity before they have accessed the relevant syntactic or semantic information. Other evidence shows the importance to sentence understanding of prosodic coherence. Prosody has also been shown to be used early in a predictive way both to overcome temporary syntactic ambiguity and to forecast the location of stressed syllables. It is hardly surprising that disruptions in prosodic structure lead to difficulties in understanding otherwise fluent sentences. If such findings can be shown to apply to spontaneous speech as well as the "lab-speech" used for most perception studies, and if disfluency results in disruption of prosodic expectations, then the detection of disfluency might be aided by prosodic events which run counter to expectations.

2.3.2 Cues suggested by previous studies

We can describe the cues suggested by previous studies in terms of the three phases in a disfluency: reparandum, editing phase and repair.

Glottalisation in the final syllable of the **reparandum** is seen as a likely cue by Bear *et al.* (1992) (and Shriberg *et al.* (1992), the SRI study) and Nakatani and Hirschberg (1993a,1993b), particularly in the case of fragments. The SRI researchers find glottalisation in 24 out of 25 vowel-final fragments. They point

out that glottalisation also occurs in fluent speech, but say that this is usually on unstressed portions of speech with low F_0, whereas, in the case of the fragments in their data, F_0 was not at the lower end of the speaker's range and usually had quite high energy. So the glottalisation they observe in repairs is thought to be acoustically distinguishable from glottalisation in fluent speech.

Nakatani and Hirschberg devote much attention to identifying fragments, because 74% of the reparanda in their data end in fragments[8]. They find that 30.2% of reparanda in their corpus have interruption glottalisation at their offsets. However 62% of the fragments in their data are not glottalised and 9% of interruption glottalisations are not in fragments. Unlike the SRI study, this one fails to take into account the fact that glottalisation is more likely to occur at a sonorant-final offset than in other phonetic environments – no distinction is made here.

Another feature of fragments which Nakatani and Hirschberg suggest may be of use in distinguishing them from full words is coarticulation. They suggest that some sonorant-final fragments "exhibit the coarticulatory effects of an unrealized subsequent phoneme" (p.49) and that when this occurs with a following pause it might be used to distinguish fragments from phrase-final words. In the example that follows (from Nakatani and Hirschberg's example (1)), the fragment *"fli-"* might, they say, be distinguishable from *"fly"* by the detection of coarticulation with a following consonant (presumably [t]).

> *"What is the earliest* **fli-** *flight from Washington to ..."*

However, it is possible to interpret this datum in a different way. If the "coarticulatory effects of an unrealized subsequent phoneme" are observed, then the speakers articulators must have moved towards the articulatory position for that phoneme. If that phoneme is a stop consonant, as may well be the case in the above example, then we can assume that a stop closure or something approaching a stop closure has occurred. In continuous speech coarticulatory and other linking behaviour between word boundaries is the norm (see, for example, Lass 1984): in many cases a word-final stop consonant would not be released until the

[8]Shriberg *et al.* find 60.2% of their repairs contain fragments, while our own corpus (36%) and Levelt's corpus (around 20%) contain a rather smaller proportion.

onset of the next word. In this particular case, if the "unrealized phoneme" is a stop consonant, then it is not so much unrealised as unreleased. Viewed this way, the word transcribed as *"fli-"* in Nakatani and Hirschberg's corpus might be better transcribed as a full word. The fact that its hypothesised final consonant is unreleased might be symptomatic not of word-interruption, but of the interruption of the phonological *link* between the last word of the reparandum and a possible fluent continuation. This topic will be followed up in more detail in Chapter 7.

Duez (1993) finds that listeners perceive interruptions in (French) speech where prepausal vowels are lengthened. Butcher (1981) (working with a corpus of read and spontaneous German speech) finds that lengthening occurs in consonants as well as vowels and that the lengthening is greater in the case of disfluent pauses than for fluent pauses. This, he contrasts with Reich's assumption (Reich, 1975) that prepausal lengthening is more likely to be longer at terminal junctures in clause and intonational boundaries than at other locations. In perceptual experiments, Butcher also finds that listeners perceive pauses within tone groups when they are only 80ms long, whereas pauses between tone groups are not perceived until they reach 220ms.

Butcher's experimental evidence is helpful in that it injects some empirical information into the controversial question of what constitutes a pause (see, for example, Rochester (1973)). Much research on speech production has taken an arbitrary time threshold for the measurement of pause. Goldman-Eisler (1958) provided other researchers with the excuse to use the time of 250msec as a standard minimum for pauses because, as she argues in 1968, shorter gaps are usually caused *"by the need to adjust the position of articulation"* (Goldman-Eisler, 1968, p.12). Other researchers have adopted other thresholds, some shorter, like Butcher, and others even up to 2 seconds or more (e.g. Siegman, 1977). Others, like Maclay and Osgood (1959), have just used subjective judgement, described as "outright useless" by Quinting (1971), who accepts Goldman-Eisler's threshold. Further evidence against Goldman-Eisler's threshold is offered by Hieke *et al.* (1983), who find that

> "Exclusion of short pauses (0.13-0.25 sec) from analysis on articulatory grounds is completely unjustified." (Hieke *et al.* (1983), p.212)

Hieke *et al.* conclude that a cut-off point of 130msec is acceptable, since although shorter pauses are detectable, as Butcher (1981) points out, they create measurement problems for both manual and automatic methods of analysis. In the latest studies on disfluency recognition, the absolute length of pause is less of an issue than the comparison of pause lengths between disfluent and fluent portions of speech.

Howell and Young (1991), who base their analysis on transcriptions from the London-Lund corpus (Svartvik & Quirk, 1980) and perform perception experiments with synthesised speech, find silent pause to be a useful cue for listeners in identifying disfluency. O'Shaughnessy (1992a, 1993b) suggests that a short pause of less than 400ms is a good indicator of oncoming repair, since this is shorter than fluent pauses, but also finds longer pauses preceding sentence restarts. Nakatani and Hirschberg (1993a, 1993b) also find that silent pauses in repairs are shorter than fluent pauses, but only significantly so where the reparandum ends in a fragment. Shriberg *et al.* (Shriberg *et al.*, 1992; Bear *et al.*, 1992) find that silent pauses can be used effectively to distinguish repairs from fluent sections of speech which have a similar syntactic pattern: repairs had a mean pause-length of 380ms, while the average gap between words with no repair ("false positives") was 42ms. One factor which most of the studies mentioned here overlook is that repairs are often performed with *no* pause at all. Blackmer and Mitton (1991) found cut-off-to-repair times of 0ms in 19.2% of overt repairs and of less than 100ms in 48.6%[9].

Hindle's (1983) editing signal is assumed to occur at the moment prior to the onset of the repair, which would be directly after the reparandum in cases where there was no overt editing phase, or at the end of the editing phase otherwise ("at the right edge of the potential expunction site" (Hindle, 1983, p.128)). He does not attempt to define the signal precisely, and no other studies have succeeded in identifying a signal – "an abrupt cut-off" of the speech signal – which is both "discrete" and "acoustically identifiable". Nakatani and Hirschberg's work (1993a, 1993b) extends the notion of editing signal to include any phenomena

[9]Howell and Young (1991) find silent pauses in only 25.5% of the repairs they examine, but their study is based on transcriptions without instrumental assistance in measuring pause durations.

which may demarcate the juncture between reparandum and repair.

Filled pauses (*um, uh*) are also potential markers of oncoming repair. But Levelt (1983) finds that in his corpus only 16.2% of overt repairs are accompanied by filled pauses. Nakatani and Hirschberg also regard filled pauses and "lexical fillers" as unreliable cues for repair, as they occur too infrequently in their data. However, filled pauses may contain prosodic information which differs according to their context and may therefore be informative where they *do* occur. O'Shaughnessy finds that filled pauses at major syntactic boundaries can be distinguished from those that occur within syntactic units by their longer duration and the longer periods of silence that surround them and by higher F_0 at onset, but he does not discuss their relationship to repairs. Shriberg and Lickley (1992a,1992b,1993) suggest that very brief filled pauses with rapidly falling F_0 often mark a repair and that an unexpectedly high F_0[10] on a filled pause may be a good indicator of a fresh start.

Lexical fillers or discourse markers, such as "I mean", "well" and "no" are not regarded as important markers of repair as they do not occur often enough (only 9.8% of repairs in Nakatani and Hirschberg's data have either lexical or non-lexical fillers). Shriberg *et al.* note, from a very small sample (9 lexical fillers at repairs and 15 in fluent speech), that instances of "well" and "no" which occur with repairs can be distinguished from the same words when used in a fluent context by simple prosodic analysis. When the words were used at the site of a repair, F_0 fell; in other contexts the words had rising F_0. The repair-marking words had no lexical stress, but the same words in fluent speech had stress. When used at a repair site, they were more likely to be accompanied by silent pause than when used in fluent speech. These findings support Hirschberg and Litman's (1987) analysis of the prosody of "now" in its different functions, which do not, however include marking repair sites: these authors found that "cue-phrase" *now* could stand alone in its own intonational phrase, whereas deictic *now* could not; they also found significant differences in the type of pitch accent that it could take depending on its rôle.

[10]Shriberg and Lickley find that the F_0 of filled pauses with no repair is related to that of their context.

Pragmatic functions of discourse markers have been described by several authors (James, 1972; James, 1973; Du Bois, 1974; Levelt, 1983; Schiffrin, 1987): Schiffrin notes that "well" is used for "background repairs", which are subordinate in the sense that they alter or add to the hearer's understanding of the surrounding speech, and "I mean" is used for "replacement repairs", which provide different information from the preceding speech, rather than supplementing it. Such markers may be seen as cues in the the sense that they can mark the onset of a repair and help the listener to understand the function of the repair, but they occur most commonly as sentence initiators, with no repair.

Studies of the aspects of the **repair** which might act as cues concentrate on prosodic features. Howell and Young (1991) find "added stress" in the repair to be a useful cue in the perception of disfluency, but, as with their findings on the use of pause in the editing phase, they rely on transcriptions, rather than analyses of the speech signal, so it is hard to assess the reliability of their data in comparison to that from instrumental studies. Cutler and Levelt (1983) find that 45% of lexical error repairs in Levelt's corpus contain prosodic marking, but their conclusion is that such marking is used by the speaker for contrastive accentuation, rather than as a specific marker of disfluency[11]. Their corpus comes from spontaneous speech in a very limited discourse domain which contains more opportunities for such contrastive stress in repairs than might be expected in a more normal setting. Speakers in Levelt's corpus are given the task of describing a pattern of coloured discs connected by arcs, as if describing a route. The majority of lexical errors are limited to errors in selection of a colour term to describe a disc and in selection of a term describing the direction to move to get to the next disc. A greater number of marked corrections were found for direction error repairs

"left to green - er, right to green"

than for colour error repairs.

"... and it ends then in a black - rather, in a purple ball"

[11]Note that the study includes only "Lexical repairs": these make up 52% of the repairs in Levelt (1983).

There were 11 colours in the patterns, whereas most direction terms were choices between two opposites "up" or "down", "left" or "right" etc..

The authors conclude that there is more intonational marking where there are fewer alternatives to choose from, but they go on to conjecture that it is probably not the number of alternatives so much as the degree of opposition between the alternatives that is the main factor underlying intonational marking. The important point to note here is that while it is commonly assumed that repairs might be intonationally marked, the evidence from Levelt and Cutler's study suggests that intonational marking in repair has a specific semantic function, which is similar to the use of contrastive stress in fluent speech, and that it is not to be seen as a sign of repair *per se*. The relatively high incidence of prosodic marking in the corpus they use (which is still a minority of the lexical repairs in any case) is likely to be due to the fact that the task that their speakers performed was very prone to provoke lexical selection errors which involved strong semantic contrasts with the corrections.

Given the limited size of the universe in Levelt's corpus and the ample opportunities for repair-marking with the type of polar contrasts which tend to attract it, it is no surprise that Nakatani and Hirschberg fail to replicate the findings in their own work on a larger and more varied corpus with generally much broader semantic fields. These authors compare F_0 and amplitude values before and after disfluent interruptions for all repair types and find a small but reliable rise both between peaks on either side of the interruption (mean rises of $4.1Hz$ and $1.5db$) and between values at reparandum offset and repair onset[12]. They find no significant differences for the same values on either side of fluent pauses, but then also find that the differences between the differences across disfluent and fluent pauses does not differ significantly and conclude that the differences in F_0 and amplitude across disfluent pauses are too small to be of use in distinguishing them from fluent pauses. O'Shaughnessy describes intonational and durational features of repairs, but only in order to distinguish one repair type from another

[12]It is also possible that the difference in criteria for judging stress between the two studies plays a rôle: Levelt and Cutler made subjective judgements of marking on the basis of perceived relative pitch, amplitude and duration of the error and correction, where Nakatani and Hirschberg used instrumental measures.

and not to distinguish disfluent from fluent prosody. Shriberg *et al.* find that F_0 can be used to distinguish repairs with the pattern $M_1|XM_1$ from potential repairs with the same pattern (e.g. "*flight* earliest flight" vs "a flight on flight number five one one"), as the peak F_0 value of X was nearly always higher than the preceding M_1 in repairs, but did not differ in the fluent versions. They report no other significant findings for F_0 values in the repair.

In summary, previous studies explore several possible avenues in the search for acoustic and prosodic cues for disfluency, but find few hard and fast rules and no universal marker. The most favoured cue is the presence of pause at the interruption site and its duration relative to either fluent pauses (where the disfluent pause is usually shorter) or sites which a pattern matcher can identify as potential repair sites (where the disfluent pause is usually longer). But the presence of disfluent pause on its own (however defined) is insufficient evidence for the presence of *repair*: pauses regularly appear in spontaneous speech with no repair, both at the end of phonological phrases, where the syntactic and intonational context mean they may be expected, and within phonological phrases, where they may be heard as hesitations; many repairs are not accompanied by pause at all. The measurement of F_0 values on either side of the interruption for the set of all repairs has yielded no useful information. Filled pauses and discourse markers are not found sufficiently frequently in repairs to be seen as strong cues. No discrete editing signal has been identified. Investigations of pause duration and F_0 and amplitude measures have looked at these features from a viewpoint strictly local to the interruption, without relating them to global structure or even local phrase structure. No work has investigated the possible rôle of prosodic *expectations* in detecting disfluency. It is, of course, likely, given the variety of possible patterns of disfluency and types of interruption, that rather than a single signal applying to all disfluencies, several different signals may alternate or combine as cues.

2.4 Goals and Methodology

Disfluency in spontaneous speech presents complex problems for any model of speech processing. Yet, everyday experience as well as some of the experimental evidence discussed above (Martin and Strange, 1968) tells us that the human

processor is able to filter out discontinuities so efficiently that they very often go unnoticed. The question of how we do it is central to this thesis. The computational models described above make the assumption that the problem is essentially one of syntactic parsing: all words are recognised in advance of parsing, and patterns of matched words or syntactic categories which are likely disfluencies can be easily identified. A model of human speech processing can make no such assumptions: in encountering a disfluency in mid-sentence, a listener will only have partial information about the syntax and is likely not to have recognised all the words in the vicinity of the interruption (Bard *et al.*, 1988). Psycholinguistic work in the area of processing disfluent speech is so scarce that fundamental questions about the perception of disfluency are unanswered. The most basic of these questions, and the first to be addressed in this thesis is:

How soon can listeners detect disfluency?

The next obvious question, and the second major question addressed here is:

What cues can listeners use in detecting disfluency?

In effect, the task is to find "recognition points" for disfluency and to match these points to the speech signal to find what information is present there.

A well-established technique for examining cues in on-line speech-perception tests is the gating paradigm (Grosjean, 1980; Cotton & Grosjean, 1984), which has usually been used in experiments examining the on-line processes involved in spoken word recognition (Bard *et al.*, 1988; Grosjean, 1980; Cotton & Grosjean, 1984; Pickett & Pollack, 1963; Pollack & Pickett, 1963; Pollack & Pickett, 1964; Tyler & Wessels, 1983; Tyler & Wessels, 1985). Gating allows a spoken stimulus to be presented to subjects repeatedly, the length of the stimulus being incremented on each presentation. Using this method it is possible to find recognition points for words to a degree of accuracy defined by the size of the gate.

For the first two experiments, word-level gating was used to look at large portions of the utterance: stimuli were presented in strings whose length was incremented by one word on each subsequent presentation, as the example in figure 2.3 illustrates. Each "gate" in this example contains all of the speech signal and any silence up to the moment prior to the onset of the next word: so

Presentation number:	**Subjects hear:**
1	it's
2	it's quite
3	it's quite obvious
4	it's quite obvious he's #
5	it's quite obvious he's # he's
6	it's quite obvious he's # he's on
7	it's quite obvious he's # he's on something

Figure 2.3. Example of sentence presented by word-level gating technique: subjects hear successive presentations of the same utterance, incremented by one word on each presentation. The symbol "#" represents a silent pause.

presentation number 4 in figure 2.3 contains the silent pause which separates the two instances of "he's". This method allowed word-by-word monitoring for cues to oncoming disfluency in a set of disfluent utterances as well as in the sets of fluent control stimuli described in Chapter 4. It also allowed the opportunity to test word recognition in the stimuli and to examine the effect of the presence of disfluency on the recognition of nearby words.

In Experiment Three, shorter (35ms) gates were used to focus the search for recognition points of disfluency on the speech immediately surrounding the interruption point and to make it possible to determine whether or not word-recognition is a prerequisite for the detection of disfluency.

One practical drawback of the gating technique is that the experiments demand a great deal of time for each stimulus. In order to keep the running time for the experiments to a length bearable for volunteer subjects, the set of stimuli had to be restricted to a fairly small number (a total of 120 stimuli, 30 of which were disfluent). An advantage of the technique which compensates for this is that each stimulus can generate a large amount of data.

In Experiments Four and Five, the focus was on the contribution of prosodic properties of the signal (pitch, intensity and rhythm) to the detection of disfluency. A method was required to treat the speech signal in such a way as to remove segmental information (and therefore lexical and syntactic information)

but leave fundamental frequency (F_0) and changes in relative amplitude and syllabic durations intact. The method chosen was low-pass filtering. Removing all spectral energy above a certain frequency level from the speech signal can have the effect of making segmental information inaudible: the average level of frequency of the first formant for the vowel [i] in English (the vowel with the lowest F_1) is around $270Hz$ for male speakers and $310Hz$ for female speakers (Denes & Pinson, 1963). The maximum F_0 levels in any of the stimuli were lower than these F_1 levels and the filter cut-off points were decided accordingly, allowing the maximum F_0 peaks to be perceived, while cutting out any segmental information (Chapter 6). The resulting signal sounds like natural but muffled speech, with audible pitch movements, amplitude differences and syllabic lengthening. An alternative method considered was spectral inversion (Blesser, 1972; Duez, 1985; Duez, 1993). Spectral inversion involves rotating the sound spectrum over a selected frequency band around a given point, such that energy peaks at low frequencies become peaks at high frequencies and *vice versa*. The resulting signal retains the original prosodic properties but, depending on the frequency band selected (200-4000Hz, for Duez), the segmental information can be corrupted. This technique was not used in the present study because the resynthesis algorithms tested did not produce a satisfactory quality of signal[13]. Another technique that has been used for examining prosody without lexical and syntactic information being available is "Reiterant Speech" (Oller, 1973; Liberman & Streeter, 1978; Larkey, 1983). This technique, otherwise and less opaquely known as "nonsense syllable mimicry", involves speakers imitating the intonation of previously prepared normal sentences, replacing each syllable with a nonsense syllable like [ma]. It is unlikely, however, that speakers could imitate disfluent speech accurately in this manner or that their products would preserve the true timing and pitch characteristics of the original speech.

One general objection which might be raised about both of the techniques used in the experiments, however, is that they do not represent realistic listening conditions.

[13]In addition, Blesser (1972) finds that speakers are able to learn to converse through spectrally-inverted speech.

In gating, the speech is presented repeatedly in ever-growing chunks, where under normal conditions the listener only hears the signal once: this repetition may give listeners more information than they would normally have in a one-pass listening situation. In spontaneous and particularly disfluent speech there may be minor disturbances in the signal which gating, particularly 35ms gating, would bring to the listener's attention more than would be the case in normal circumstances. There is much evidence from the phonemic restoration illusion (e.g. Warren, 1984; Samuel and Ressler, 1986; Tougas and Bregman, 1990; Repp, 1992) that not all phonetic events in speech are relevant to the eventual interpretation of the message: but in 35ms gating, the type of disturbance in the signal introduced in studies of the phonemic restoration illusion (phonemes obliterated or replaced by noise) would be quite likely to be brought to the attention of the listener. Under different levels of attention, the relative importance to the listener of different features of the speech signal has been found to vary: Gordon *et al.* (1993) found that the importance of voice onset time to the identification of voiced and voiceless stop consonants and the importance of formant pattern to the correct identification of vowels [i] and [I] was lessened under a low-attention condition, while the importance of F_0 onset frequency and duration for the respective distinctions increased. According to the evidence in Martin and Strange's work (1958), many disfluencies may be missed or misplaced under more realistic conditions. Gating stimuli are, if anything, more likely than single-pass presentations to force listeners to notice disfluencies and place them correctly.

The first objection to the gating technique, that repeated exposure to the same speech might have a facilitatory effect on perception, has been addressed by Cotton and Grosjean (1984) and Bard *et al.* (1988) , who show that single presentations of a string, incremental presentations of a string and repeated presentations of a string yield the same perceptual accuracy.

The second objection has more validity. But in these experiments, the aim is not to discover what cues listeners *do* respond to in processing disfluent speech under normal circumstances, but rather to use the human speech processing mechanism as a sophisticated speech processing tool, to find points where in the signal there is enough information available for a decision to be made about the fluency of the stimulus and then to establish what cues the listener has responded to. For

this reason, the speech was recorded and played back under optimal conditions, rather than attempting to reproduce suboptimal but "normal" conditions.

The low-pass filtering technique also creates artificial listening conditions: it is not normal for speech to suddenly become incomprehensible mid-utterance as if the speaker's face has suddenly been covered by a pillow. Furthermore, there is no independent evidence that listeners *are* able to make judgements about the prosody of speech when it is low-pass filtered: to the author's knowledge, it is a technique previously untried in psycholinguistic experimentation. But the experiments compare responses to disfluent and fluent stimuli: any positive difference found tends to argue that sufficient prosodic information for some discriminations is retained in low-pass filtering.

2.5 Conclusion

This chapter describes the context for the rest of the study. A set of types of disfluency has been described and terminology defined for use in the chapters that follow. Previous studies which approach the problem of processing disfluent speech come chiefly from the field of Computational Linguistics. In CL it has usually been assumed that the problem is mainly one of parsing "ill-formed input", which can be marked as ill-formed because it is rejected by an initial parse. Other approaches look for a special signal or certain patterns in the speech which trigger mechanisms for the identification and removal of the reparandum. No approaches to the understanding of disfluent speech see speech processing as an on-line incremental task. In this study, we assume such an approach and make use of the gating technique to analyse subjects' responses to disfluency as it unfolds. Low-pass filtered speech will be used to examine the effect of disfluency on the perception of prosody. The human subject is viewed as a sophisticated tool for speech understanding in the experiments, used to find cues to disfluency at the earliest possible points in the speech signal: this process is not claimed to represent what happens in natural listening conditions, where it seems that disfluency may often pass unnoticed, but it allows us to locate and subsequently identify potential cues . Analyses of the prosodic and acoustic properties of disfluent speech for the purposes of understanding speech have compared events

around interruption points with events at other specific sites – potential repairs or phrase boundaries. In this study comparisons will be made with points which are structurally and prosodically similar in fluent speech. The acoustic analyses of the stimuli, which follow the experiments, match subjects' responses to the signal to show how features of the signal affected judgements of disfluency and suggest what acoustic and prosodic features are used as cues.

Chapter 3

The Corpus

In this chapter we discuss the corpus of spontaneous speech collected to provide data for the analyses of disfluency and stimuli for the experiments. The chapter falls into three main sections: in the first section, we describe how the corpus was gathered, recorded and transcribed and introduce the typology used for describing the data; the second section looks at the frequency and distribution of the types of disfluency identified in the corpus; the third section concludes the chapter by describing the choice of stimuli for use in the experiments and the acoustic and prosodic analyses described in the chapters that follow.

3.1 Method

3.1.1 Recording

The first practical step in this project was to gather a corpus of natural, spontaneous speech of sufficient size to provide the raw material for textual and acoustic analyses and stimuli for perception experiments. Two major areas of concern affected the choice of method for gathering the speech data required: the technical quality of the recordings and the content and style of the speech.

A high quality of recording was required to provide the optimal materials for experimental stimuli and for accurate acoustic and prosodic analysis. This necessitated the use of a sound-proofed recording studio and a digital recording system. Using a studio also allowed control of the recording conditions for all

recording sessions so that there would not be qualitative differences between the recordings for the different speakers whose contributions made up the corpus. Another factor affecting the quality of the speech signal in normal conversation is interruption and overlapping speech from other speakers. In order to keep this interference to a minimum, the conversations were recorded between only two participants: one of the participants was the author, who deliberately avoided interrupting and overlapping as far as possible.

The type of interaction in which speech takes place can control various aspects of its style and content. Maclay and Osgood's (1959) analysis of the distribution of disfluencies uses as data only utterances of 80 words or more by participants at a conference. Blankenship and Kay (1964) take their data from a set of public speeches. In Levelt's corpus (Levelt, 1983), speech was in the form of a monologue and the language limited to descriptions of a pattern of coloured circles linked by straight lines. A more recent corpus which has been used for the description and analysis of disfluency involves spoken interaction in a less restricted domain, but still one which limits the speakers' syntax, and deprives them of a human interlocutor: Bear *et al.*, Shriberg *et al.* (1992), O'Shaughnessy (1992a, 1992b, 1993a, 1993b) and Nakatani and Hirschberg (1993a, 1993b) have all made use of speech from the ARPA Airline Travel and Information System (ATIS) corpus (MADCOW, 1992), in which speakers ask for information about air travel from a computerised database. In both Levelt's and the ARPA corpus the discourse context limits the range of vocabulary and syntactic structures likely to occur and puts the speaker in the unusual situation for spontaneous speech of talking to a machine (a tape recorder in Levelt's case). Another recent corpus of spontaneous speech which is being used for work connected with disfluency (Carletta *et al.*, 1993b; Carletta *et al.*, 1993a), the HCRC Map Task Corpus (Anderson *et al.*, 1991), consists of 128 task-orientated dialogues. Speakers are given different versions of a simple map with around fourteen landmarks on it. The versions have a different (intersecting) set of landmarks, and one version has a route marked. The speakers' task is to collaborate in filling in the route on the other map, without comparing the two visually. Conditions are varied for inter-speaker familiarity, eye-contact, and familiarity with the particular map (the same speakers take part in a number of dialogues).

In building up the present corpus, it was decided to allow a much freer discourse domain and to provide the natural scenario of an informal conversation between two people who were familiar with each other.

Recording sessions took place in a professional-standard recording studio. A Sennheiser MKH 815 T gun microphone was used and the conversation was digitally recorded through a Soundcraft 200B mixing console (used to control the recording level only) via a Sony PCM 701 ES digital audio processor onto the videotrack of a Betamax video cassette recorder with 14-bit resolution with emphasis at a sample frequency of 44.1KHz [1]. The six recordings each lasted between 35 and 45 minutes.

Six informants, three male and three female, aged between 25 and 45, each volunteered to take part in one recording session. All were friends or acquaintances of the author. All spoke with accents which did not differ greatly from standard British English and which subjects who took part in the experiments which employed the corpus could be expected to understand.

For the recording session, the informant and the author were seated at a table in a recording studio. The informant had been told nothing about the purpose of the recording and was simply invited to take part in a normal conversation over a cup of coffee. Since it was impossible to hide the fact that the speech produced by the informant was more important than that produced by the author, the seating positions were such that the microphone was closer to the informant, but the informant was told not to speak into the microphone, but just to ignore its presence as far as possible. The author participated normally in the conversation, but tried to encourage the informant to take the greater part. In addition, the author tried to make sure that the topic of conversation changed from time to time in order that a wide range of vocabulary would be used and that the informant would be exposed to topics in which they had various levels of interest or personal experience.

[1] This was a digital *audio* recording onto video tape – visual information was *not* recorded.

3.1.2 Transcription

Method

All the conversations were copied onto audiocassettes and transcribed by the author, using a Marantz Superscope C205 cassette tape player, with speed control, and Revox 3100 semi-open stereo headphones. For the sake of accuracy, the transcription required frequent passes over short sections of speech (frequent use of the "recall" button) and several full passes. The initial transcriptions were completed after at least four full passes; subsequent complete passes were made at various points in the development of the thesis. Any sections which were found to be too difficult for transcription by ear alone (usually disfluent sections) were examined using ILS or ESPS/Waves+ signal processing software.

Conventions

Standard orthography was used to transcribe most words in the speech except for part-words (usually at the interruption point in a disfluency): these were transcribed using phonetic symbols from MRPA (Machine Readable Phonetic Alphabet, designed at the Centre for Speech Technology Research, University of Edinburgh) and enclosed in square brackets. Special symbols marked non-speech sounds such as laughter and coughing, overlapping speech and editorial comments. No punctuation was used in the transcriptions and individual sentences were not marked.

A system of symbols was devised for marking disfluencies. In categorisation, a fundamental division was made between disfluencies which just consisted of some form of pause and those which involved some retracing or repetition.

The pause types were classed as pausal lengthening, silent pause, filled pause (usually *"um"* or *"uh"*), other non-lexical vocal sounds (*breath, creak and unintelligible sounds*) and lexical fillers (e.g. *"I mean"*, *"well"*, *"sort of"*). Pausal lengthening and silent pause were marked on the basis of the subjective judgement of the author, with no physical measurement of the durations of the lengthened segments or of the pauses: this is clearly an unsatisfactory system, since the perception of pauses and lengthening in continuous speech is a notoriously difficult task (Butcher, 1981; Duez, 1993). The analysis of the resulting data is not,

however, intended to be any more than very informal.

The inclusion of lexical fillers in a count of disfluencies is very debatable, since they do not themselves form disfluencies in the same way as pauses and repairs, but appear in specific contexts, usually not at repair sites (see Section 2.3.2, page 41). But since they are often associated with disfluency in the literature, all cases of standard lexical fillers were marked.

Both repetitions and false starts were subcategorised according to the length of the reparandum. Length was specified in an *ad hoc* manner, dividing the data into 6 categories which allowed identification of repairs by the number of words that formed the reparandum and by the presence of fragments at the end of the reparandum. The length categories differentiated between reparanda of less than one word (a fragment), exactly one word, one word plus a fragment, exactly two words, two or more words plus a fragment and more than two whole words. Further division into smaller categories would, of course, be possible, but this was not necessary for our purposes. Many different possibilities for categorisation on the basis of length of reparandum with different units of length (syllables, feet etc.) would be possible, and possibly more appropriate, but this choice was made at an early stage in the study and in the absence of any previous classification along the same lines, in order to keep the task of labelling the data as straightforward as possible.

In addition to the length subcategorisation, repetitions were marked for the number of repeats. False starts were also subcategorised further, distinguishing between cases where the informant restarted with a new sentence, restarted changing a word, restarted adding or removing a qualifier or restarted changing the pronunciation of something in the reparandum.

3.1.3 Textual analysis

Word counts and disfluency counts were computed using combinations of standard UNIX commands and shell scripts. Speech by the author was excluded from all the analyses.

3.2 Distribution of disfluencies

This section describes the numerical distribution of disfluencies, editing terms and lexical fillers in the corpus. The motivation of this study is to demonstrate the great frequency with which spontaneous conversational speech is interrupted by disfluency and to provide the background for the selection of experimental stimuli. Although several other studies have provided similar surveys of the distribution of disfluencies (e.g. Maclay and Osgood (1958), Blankenship and Kay (1964), Allwood *et al.* (1983), Levelt (1983), Blackmer and Mitton (1991)) it is outwith the scope of this thesis to make detailed comparisons here: and, as was seen in Chapter 2, there are almost as many coding schemes as there are surveys of distribution, so it difficult to compare like with like.

Six dialogues of between 35 and 45 minutes duration were transcribed as described above.

Conversation was allowed to flow freely, but the author made sure that the topic changed in a natural way at various points. Topics covered included education, work, sport, alcohol and politics, and were approached from the viewpoints of personal experience as well as abstract discussion.

The transcriptions yielded a total of 22,767 words, excluding fragments (491) and non-lexical pause-fillers (705) but including all complete words in reparanda (1562) and all lexical fillers (528 tokens, yielding 919 words). A breakdown of the word totals by informants is shown in table 3.1: the range of numbers of words per informant reflects different ratios of informant talking time to experimenter talking time as well as different lengths of conversation.

The distribution by informants, and frequency in words per token, of pause-types and lexical fillers is shown in table 3.2. The frequency is the ratio of the total number of words (taken from the bottom row of table 3.1) and the number of tokens of the type of pause or filler. It is clear from table 3.2 that the informants differed considerably in the frequency with which they used pause devices and lexical fillers. Further analyses tested whether the choice of filler expressions also differed between informants.

Filled pauses were realised in two ways in the data, transcribed as *"um"* and *"uh"*: the relative frequencies of *"um"* (75.9% of all filled pauses) and *"uh"*

Count	Informants G	H	J	M	N	P	Total
"Fluent" words:	4306	2795	3101	3002	3531	3751	20486
%	*92.4*	*93.3*	*87.0*	*91.9*	*83.0*	*88.8*	*89.2*
Words in lexical fillers:	119	53	165	107	348	127	919
%	*2.5*	*1.8*	*4.6*	*3.3*	*8.2*	*3.0*	*4.0*
Words in reparanda:	235	147	299	159	376	346	1562
%	*5.0*	*4.9*	*8.4*	*4.9*	*8.8*	*8.2*	*8.2*
Total	4660	2995	3565	3268	4255	4224	22967

Table 3.1. Corpus analysis: Word counts by informant.

Type	Informants G	H	J	M	N	P	Total
Silent pause	100	53	86	122	90	26	477
f	*46.6*	*56.6*	*41.4*	*26.8*	*47.3*	*162.5*	*48.1*
Filled pause	31	30	152	79	124	289	705
f	*150.3*	*99.8*	*23.5*	*41.4*	*34.3*	*14.6*	*32.6*
Other	30	6	13	15	27	34	125
f	*155.3*	*499.2*	*274.2*	*217.9*	*157.6*	*124.23*	*183.7*
Lexical Fillers	69	38	94	65	193	69	528
f	*67.5*	*78.8*	*37.9*	*50.3*	*22.0*	*61.2*	*43.5*

Table 3.2. Corpus analysis: Distribution by informants and frequency (*f*) in words per token of pauses, filled pauses (*um, uh*), other non-lexical fillers (lengthening, breath, creak and unintelligible sounds) and lexical fillers.

(24.1% of all filled pauses) did not differ significantly between informants.

Eight types of lexical fillers were found (table 3.3): *"I mean"*, *"well"* and *"you know"* were present in the speech of all informants and all informants used one or both of *"sort of"* and *"kind of"*: these 5 types accounted for 98.5% of all lexical fillers. The remaining 3 types were only used by one (*"sorry"*) or two (*"like"* and *"oh"*) informants. Informants differed considerably in the relative frequency with which they used the three most common lexical fillers: the difference was highly significant ($\chi^2 = 62.91$, $df = 10$, $p < 0.0001$).

One overt editing sentence was found in the whole corpus: *"no, sorry, scrap that"*.

Over the whole corpus, speech was interrupted by a repetition or a false start every 20 words. These interruptions consisted of 624 repetitions (one every 36.8 words) and 522 false starts (one every 44 words). As table 3.4 shows, there were considerable differences in fluency between informants.

The most frequent type of repetition was the single-word repetition ($N = 317$), followed by part-word repetitions ($N = 171$), two-word repetitions ($N = 67$) and repetitions of a single word plus a fragment ($N = 44$): all informants produced examples of all of these types. Repetitions of more than two whole words were less common ($N = 17$) and repetitions of two or more words plus a fragment were the least common type ($N = 8$). Double repetitions (*"if if if their view ..."*) formed 12% of all single word repetitions and at least two were produced by each informant; 24.6% of all single fragment repetitions were double repetitions, but 86% of these were from one informant, and only two other informants produced any tokens. Double repetitions of other types were very rare, as were triple, quadruple and quintuple repetitions (table 3.5)[2]

The vast majority of single-word repetitions were of function words: only 13 repetitions of content words were found, 4.1% of the total of single-word repetitions. This distribution differs significantly from that expected on the basis of the frequency of function and content words in the data, as estimated from a random sample of 900 words in which 33% of words were content words ($\chi^2 = 139$, $df = 1$, $p < 0.0001$). There was thus a clear tendency for function

[2] "Whole single words" in this study include elisions *"it's"*, *"I've"*, *"you're"* etc..

Type	Informants G	H	J	M	N	P	Total
"I mean"	12	3	32	20	69	24	160
%	*17.4*	*7.9*	*34.0*	*30.8*	*35.7*	*34.8*	*30.3*
"well"	18	23	23	20	36	9	129
%	*26.1*	*60.5*	*24.5*	*30.8*	*18.6*	*13.0*	*24.4*
"you know"	33	8	26	10	24	23	124
%	*47.8*	*21.0*	*27.7*	*15.4*	*12.4*	*33.3*	*23.5*
"sort of"	5	0	13	1	29	10	58
%	*7.2*	*0*	*13.8*	*1.5*	*15.0*	*14.5*	*11.0*
"kind of"	0	4	0	11	33	1	49
%	*0*	*10.5*	*0*	*16.9*	*17.1*	*1.4*	*9.3*
"like"	0	0	0	3	1	0	4
%	*0*	*0*	*0*	*4.6*	*0.5*	*0*	*0.8*
"oh"	1	0	0	0	1	0	2
%	*1.4*	*0*	*0*	*0*	*0.5*	*0*	*0.4*
"sorry"	0	0	0	0	0	2	2
%	*0*	*0*	*0*	*0*	*0*	*2.9*	*0.4*
Totals	69	38	94	65	193	69	528

Table 3.3. Corpus analysis: Distribution by informants of lexical fillers.

Type	Informants G	H	J	M	N	P	All
Repetitions	63.0	65.1	26.3	48.8	38.0	21.4	36.8
False Starts	60.5	68.1	33.3	65.4	34.6	33.3	44
All	30.9	33.3	14.7	27.9	18.1	13.0	20.0

Table 3.4. Corpus analysis: Rate (number of words divided by number of disfluencies) of repetitions and false starts, by informants.

Type	Number of repeats 1	2	3	4	5	Total
Fragment	120	42	8	0	1	171
1 Whole Word	273	38	5	1	0	317
1 Word + Fragment	42	2	0	0	0	44
2 Whole Words	66	1	0	0	0	67
$\geq$ 2 Words + Fragment	8	0	0	0	0	8
$>$ 2 Whole Words	17	0	0	0	0	17
Total	526	83	13	1	1	624

Table 3.5. Corpus analysis: repetitions by type and number of repeats.

Clause Position	Word Class		
	Function	Content	All
Initial	208	2	210
%	*99*	*1*	*100*
Medial	96	11	107
%	*89.7*	*10.3*	*100*
Total	304	13	317
%	*95.9*	*4.1*	*100*

Table 3.6. Corpus analysis: Distribution of single-word repetitions by word class and clause position.

words to be repeated rather than content words. The majority (66.2%) of single-word repetitions were at the onset of a major syntactic clause. Most (68.4%) repeated function words were clause-initial but most (84.6%) repeated content words were clause-internal (Table 3.6). The relative distribution of function and content words in clause-initial position (99% and 1%, respectively) did not differ significantly from that expected on the basis of the overall frequency of word classes at clause-initial position: a sample of 240 randomly selected clause-initial words showed that 96.7% were function words and 3.3% content words.

A total of 223 repetitions (35.74%) had reparanda ending in a fragment. The intended word in single-fragment repetitions was more likely to be a content word than would be predicted by the overall distribution of word classes, both in clause-medial and clause-initial position. In clause-medial position, the majority of repeated fragments were of intended content words (80.9%). In clause-initial position, 10.3% of repeated fragments were of content words, significantly more than expected ($\chi^2 = 4.11$, $df = 1$, $p < 0.05$). Overall, there were more intended function words than intended content words in single-fragment repetitions (55% and 45%, respectively), but this showed a significant tendency for speakers to interrupt content words, rather than function words, given the overall distribution

Clause Position	Intended Word Class		
	Function	Content	All
Initial	78	9	87
%	*89.7*	*10.3*	*100*
Medial	16	68	84
%	*19.1*	*80.9*	*100*
Total	94	77	171
%	*55*	*45*	*100*

Table 3.7. Corpus analysis: Distribution of single-fragment repetitions by class of intended word and clause position.

of the classes (67% function words and 33% content words) ($\chi^2 = 9.21$, $df = 1$, $p < 0.01$).

False starts were divided into four categories: pronunciation changes; addition or deletion of qualification; word substitution; complete change of sentence. Sentence changes were the most common type, making up more than a half of the total number of false starts (53.2%); word substitutions were the second most frequent (25.3%), followed by qualification changes (13.8%) and pronunciation changes (7.7%). Two informants showed idiosyncratic differences from the others in the distribution of types of false start ($\chi^2 = 28.89$, $df = 15$, $p = 0.0166$) (Table 3.8): informant J stopped to qualify his original utterance more often than other informants; informant P changed pronunciation more often than others (usually correcting tongue-slips).

A total of 522 false starts (37.36%) had fragment-final reparanda. The incidence of fragment-final reparanda varied with the type of false start. A large proportion of word substitutions (56%) and pronunciation changes (65%) involved word-interruptions (fragment-final reparanda). The interruptions in sentence restarts and qualifications were usually after full words (76.3% and 62.5%,

Change Type	Informants G	H	J	M	N	P	Total
Pronunciation	3	3	3	4	9	18	40
%	*3.9*	*6.8*	*3.0*	*8.0*	*7.3*	*14.2*	*7.7*
Qualification	11	2	24	4	14	17	72
%	*14.3*	*4.5*	*23.8*	*8.0*	*1.4*	*3.4*	*13.8*
Word	17	12	26	12	29	36	132
%	*22.1*	*27.3*	*25.7*	*24.0*	*23.6*	*28.3*	*25.3*
Sentence	46	27	48	30	71	56	278
%	*59.7*	*61.4*	*47.5*	*60.0*	*57.7*	*44.1*	*53.3*
Total	77	44	101	50	123	127	522
%	*100*	*100*	*100*	*100*	*100*	*100*	*100*

Table 3.8. Corpus analysis: Distribution of false starts by informants.

Change Type	Reparandum Length 1	2	3	4	5	6	Total
Pronunciation	23	5	6	4	1	1	40
%	*57.5*	*12.5*	*15.0*	*10.0*	*2.5*	*2.5*	*100*
Qualification	17	20	6	12	4	13	72
%	*23.6*	*27.8*	*8.3*	*16.7*	*5.6*	*18.1*	*100*
Word	51	33	9	11	12	16	132
%	*38.6*	*25.0*	*6.8*	*8.3*	*9.1*	*12.1*	*100*
Sentence	21	72	20	61	25	79	278
%	*7.5*	*25.9*	*7.2*	*21.9*	*9.0*	*28.4*	*100*
Total	112	130	41	88	42	109	522
%	*21.5*	*24.9*	*7.8*	*16.9*	*8.0*	*20.9*	*100*

Table 3.9. Corpus analysis: Distribution of false starts by length of reparandum: length 1 = single fragment; length 2 = single word; length 3 = single word plus fragment; length 4 = two whole words; length 5 = two or more words plus a fragment; length 6 = three or more whole words.

respectively). Levelt (1983) finds a pragmatic explanation for speakers interrupting words:

> "Interrupting a word signals that that word is wrong" (Levelt, 1983: p.63).

The above results suggest a trend in the direction of Levelt's conclusion, but the different coding schemes make an accurate comparison difficult.

The length of reparandum also varied with the type of false start. Sentence restarts had reparanda of 2 words or more in 59.3% of cases, whereas the reparanda for the other three types were less than two words in length in the majority of cases (Table 3.9).

Clause Position	Intended Word Class		
	Function	Content	All
Initial	23	3	26
%	*88.5*	*11.5*	*100*
Medial	16	58	74
%	*21.6*	*78.4*	*100*
Total	39	61	100
%	*39*	*61*	*100*

Table 3.10. Corpus analysis: Distribution of single-fragment false starts by class of intended word and clause position.

For false starts with reparanda consisting of single fragments, it was often possible to identify the intended word or, failing that, the lexical class of the intended word, via contextual and syntactic cues. As was the case with repetitions, false starts whose reparandum consisted only of a fragment were usually (61%) intended content words. In most cases, single fragments in false starts were clause-medial (74%). The tendency for content words, rather than function words, to be interrupted was apparent both in clause-initial and clause-medial positions (Table 3.10). In 12 of the 112 cases examined, it was impossible to decide the lexical class of the intended word.

Fragments in reparanda longer than one word were also more frequently content words than function words: combining false starts and repetitions for lengths "3" and "5" ($N = 135$), 10 cases were excluded because the intended word could not be guessed and of the remaining 125, 99 fragments (79.2%) were intended content words and 26 (20.8%) function words.

Filled pauses and lexical fillers did not typically mark the interruption in repetitions and false starts. The interruption was accompanied by a filled pause in only 75 cases (6.54%) of a total of 1146 repairs (where repairs are all repeats and all false starts); this figure represents 10.64% of all filled pauses. While

"um" predominated in filled pauses where there was no repair (84.9%), *"um"* and *"uh"* were almost equally frequent at repair sites (52% and 48%, respectively), although *"uh"* was only used by three informants at the relatively few (75) repair sites where a filled pause marked the interruption.

Lexical fillers accompanied the interruption in 60 cases (5.24%); only 11.36% of lexical fillers were at disfluent interruptions.[3] Of the 8 types of lexical fillers identified, 5 were found at the interruption point in disfluent interruptions. One (*"sorry"*, $(N = 2)$) was exclusively used in repairs and appeared at points where the speaker wished to alter something in the original utterance. The others usually occurred in the same syntactic positions with respect to the repair as they did in otherwise fluent speech: "I mean" and "well" are typically sentence-initiators in fluent spontaneous speech; "you know" is also a sentence initiator, but also occurs frequently at the end of sentences; "sort of" is used in the position of a qualifier, before content words. In the repairs, "I mean" was used at the interruption point in 24 false starts and 3 repetitions, all with sentence-initial restarts; "well" was used at the interruption in 11 false starts and 4 repetitions and the repair was sentence-initial in all but one cases; "you know" was used at the interruption in 10 false starts and 4 repetitions, all with sentence-initial repairs; "sort of" was used with one sentence repair and two fragment repetitions, where the filler seemed to act as an inserted qualifier ("l@- sort of looney"). The low frequency of lexical fillers at repair sites and their relatively greater frequency in fluent contexts leads us to the conclusion that they should not be viewed as disfluency markers, but that where they appear at the onset of repairs they are just playing their normal rôles in the discourse, as defined, for example, by Schiffrin (1987)).

3.2.1 Summary

In summary, disfluent phenomena were very frequent in the corpus, with some form of disfluency (including lexical fillers) occurring every 7.7 words over all (every 9.4 words, excluding lexical fillers).

[3]Of the 477 silent pauses perceived by the transcriber, 55 (11.53%) were at repair sites (5.26% of repairs), but this data is probably not reliable, for the reasons explained in section 3.1.2.

There was considerable inter-informant variation in the frequency of disfluency: for example, the frequency of repairs varied from one every 13 words to one every 33.3 words (table 3.4). It is interesting to note that, for our small number of informants, female speech was less disfluent than male speech in every case (females are informants G, H and M). There was also some idiosyncratic inter-informant variation in the style of disfluency and use of lexical fillers.

The analysis of the data was not intended to be a fully comprehensive survey of the distribution and syntax of disfluencies in the corpus: such a study was beyond the scope of this thesis. In addition to the simple numerical distribution of pause phenomena, repetitions and false starts we examined the position and lexical class of fragments, the lexical class of single-word repetitions and the frequency of occurrence of fillers at the interruption point for repairs.

Interrupted words are more likely to be content words than predicted by chance both clause-initially and clause-internally and both in repetitions and in false starts.

Single-word repetitions are much more commonly function words than content words.

The word following (or forming) an interruption for repair is only infrequently a filler in this corpus. Only 11.78% of repairs were accompanied by a filler at the interruption.

3.3 Selection of Experimental Stimuli

The perception experiments planned required a selection of repairs from the corpus, an equal number from each informant, which would reflect the relative frequency of the various types identified as far as possible. The gating technique has an unfortunate drawback as far as the amount of disfluent data to be presented is concerned: the repeated presentation of gradually incrementing stimuli takes a great deal of time per stimulus. Human subjects, however, can not be expected to donate more than a certain length of time to psycholinguistic research. In order to keep the length of the experiments down to what was thought to be an acceptable time (maximum 1h 45m, with breaks), the number of disfluent stimuli was restricted to 30, 5 from each informant, with a total of 90 control stimuli.

The first division of the planned stimuli was into repetitions and false starts. It was decided to represent these two categories equally in the stimulus set (15 repetitions and 15 false starts), even though an overall count shows that repetitions represent a larger proportion (54.5%) of the set of all repairs. Within each of these categories, the next division was by the length of the reparandum. Within each repair category (repetitions and false starts), the proportion of repairs of each length in the corpus as a whole was calculated, multiplied by 15 and rounded to the nearest integer to determine how many stimuli with that length of reparandum would be included. For example, there were 317 single words repetitions, 0.508 of the total of 624 repetitions. A comparable proportion of 15 gives about 8, which is the number of single-word repetitions represented in the stimuli.

Each informant was represented in the stimulus set by 5 repairs: these were divided between the repetition and false start sets so that 3 informants provided 2 repetitions and 3 false starts and the other 3 informants provided 3 repetitions and 2 false starts. As far as the distribution allowed, the choice of which subcategory of repair was provided by which informant depended on how heavily a subcategory was represented in the set of repairs for an informant. The resulting distribution pattern of repairs used as stimuli is shown in table 3.11.

Having defined the categories from which the disfluent stimuli were to be taken, three other factors were taken into account in selecting stimuli. First, turn-initial repairs with a reparandum of a single word or shorter were avoided, in order to allow at least one word of immediate left context in the stimuli. Second, as all stimuli were to be presented with about ten seconds of the previous discourse context, it was important that that context should be reasonably clear and not itself confused by too much disfluency. Thirdly, it was important that the stimulus itself should not contain extraneous noise or overlap by the other speaker, which could accidentally influence judgements in the experiments.

Following the above criteria, 30 disfluent utterances were selected from the corpus to be used as stimuli in the experiments.

Next, another 30 utterances were chosen from the corpus to provide spontaneous fluent controls for the disfluent stimuli, each member of a disfluent-fluent pair coming from the same informant. These items were selected to match the

Type of Disfluency	Code	Number of Cases	Distribution across speakers G	H	J	M	N	P
Repetitions:	R	15						
Fragment	R1	3	0	1	1	0	0	1
One word	R2	8	2	2	1	1	1	1
One word and a fragment	R3	1	0	0	0	0	0	1
Two words	R4	2	0	0	0	1	1	0
More than two words	R6	1	0	0	1	0	0	0
False starts:	C	15						
Fragment	C1	3	1	0	0	0	2	0
One word	C2	4	0	1	0	1	0	2
One word and a fragment	C3	1	0	0	0	1	0	0
Two words	C4	3	2	0	1	0	0	0
Two words and a fragment	C5	1	0	0	0	0	1	0
More than two words	C6	3	0	1	1	1	0	0
Totals		30	5	5	5	5	5	5

Table 3.11. Corpus: Distribution of disfluency types used as stimuli.

beginning (Original Utterance) of the disfluent stimuli for structure, length and prosody as far as possible. Finding closely matched pairs of utterances in a corpus of between 3000 and 4700 words per informant in free conversation is not an easy task. The initial search was made from the transcriptions on the basis of matching strings and structures. Where that failed, looser matches were sought and subjective judgements of prosodic similarity were made by the author by listening to the recordings. The average length of the fluent utterances thus selected was .833 words shorter than the disfluent utterances.

To provide stringent controls for the disfluent test items, the ideal would be fluent versions of the same utterances. This being impossible in a corpus of spontaneous speech, a method for providing "the next best thing" was devised: each spontaneous item used in the experiment was matched with a fluent rehearsed version produced in the following way. Each disfluent test item was edited using ILS on MASSCOMP, to produce, where possible, a fluent-sounding version of the original utterance. Where it was impossible to produce a fluent-sounding version of a disfluent utterance, (in 7 cases) the original utterance up to the interruption point was recorded. The resulting utterances were recorded onto an audiocassette, mixed in random order with the original fluent test items, each item being repeated six times. The original speakers were then asked to listen to their section of the tape and to repeat what they heard as accurately as possible. A script was provided as an aid, but the speakers were encouraged to imitate, rather than read. The only occasion where the script had to be used was where it had been impossible to produce a full utterance from the original disfluent utterance: on these occasions, the speaker was asked to complete the utterance by reading the continuation suggested on the script. The speakers' responses were recorded in the same studio and 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 as determined aurally by the author. An example of one of the resulting sets of stimuli is given below (Examples 3.1, 3.2 and 3.3).

Example 3.1 : Spontaneous Disfluent:

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

Example 3.2 : Rehearsed "Disfluent":

it's quite obvious he's on something

Example 3.3 : Spontaneous and Rehearsed Fluent:

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

Example 3.3 illustrates the type of spontaneous control that was selected. In this case, the closest match that could be found for the OU *"it's quite obvious he's"* is *"we know that it's"*. The two sentence onsets have similar syntactic structures, speech rates and stress and intonation patterns.

All stimuli used in all the experiments are listed in Appendix A.

Chapter 4

Word-level Gating Experiments

In this Chapter, two word-level gating experiments are described which address the question of *how soon* disfluency can be detected in the on-line processing of speech. The first experiment tests for cues *before the onset* of repair, particularly testing the hypothesis that listeners perceive an editing signal at the moment before the repair begins. In the second experiment, subjects are asked to note when they perceive that disfluency *has* occurred. In addition to the disfluency detection tasks, in both experiments subjects perform word recognition tasks at each gate. This allows us to test for possible effects of the presence of disfluency on word recognition and for effects of non-recognition of words on the detection of disfluency.

4.1 Experiment 1: Finding Oncoming Disfluency

Since there is at time of writing no direct experimental evidence to support claims about how the human speech processing mechanism handles disfluent speech, we can only make casual observations based on everyday experience and anecdotal and indirectly relevant empirical evidence. Everyday experience of listening to conversation, which is typically peppered with disfluency, suggests that the HSPM can interpret such speech with great efficiency. Just as utterances which are on paper "garden path" sentences do not usually trouble the listener for contextual (Altmann, 1985; Crain & Steedman, 1985; Paul *et al.*, 1992) or prosodic

reasons (Beach, 1991; Marslen-Wilson *et al.*, 1992), it seems that local ambiguities and potential parsing problems inherent in transcriptions of normal disfluent speech are handled so smoothly that they go virtually unnoticed by listeners: even when asked to pay close attention to disfluencies in a listening and production task, Martin and Strange's subjects had great difficulty in identifying, placing and correctly reproducing them (Martin & Strange, 1968). The apparent speed and efficiency with which the HSPM can process speech with disfluency suggest that rather than depending on the resolution of parsing problems which might only become apparent much later in the utterance, the problem of detecting disfluency might actually be solved very early.

Hindle (1983) suggests that a cue to help listeners detect disfluency, might be a *discrete* "editing signal", specifically

> ... a phonetically identifiable signal placed at the right edge of the potential expunction site ... *(Hindle, 1983, page 128).*

If such a signal *is* present in disfluent speech, then this suggests that listeners have a valuable early cue which they are able to use in solving the recognition problem. If, as Hindle suggests, the signal is at the end of the original utterance (to translate "expunction site" to our chosen terminology) then this suggests that, given all the speech up to the end of the original utterance, a listener should be able to detect this signal and be immediately prepared for the onset of a disfluency. The success of Hindle's algorithm depends on the detection of such signals, apparently independently of word recognition and initially of parsing, too: the presence of repair is confirmed by the discovery of repeated parts of text or certain grammatical configurations (see Chapter 2, Section 2.2.2).

The experiments in this thesis were designed to examine the points in the speech signal at which human listeners are able to detect disfluency. The first experiment looks at the possibility suggested by Hindle's notion of a discrete editing signal – that listeners will be able to detect oncoming disfluency because of the presence of such a signal at the very end of the original utterance. Aside from the editing signal notion, it is possible that other cues contained in the original utterance may warn listeners that disfluency is about to occur: silent pause, lengthening at the end of the reparandum and glottalisation have been

suggested as possible signals, but no published work has found consistent and reliable cues of this sort (see Chapter 2, Section 2.3.2). One other possibility in this experiment is that listeners will react to their inability to recognise a word at first presentation by indicating that they have detected oncoming disfluency.

The effect on the recognition of words of the presence of an interruption is also of interest in the on-line processing of disfluent speech. Given that a certain percentage of words in fluent speech are recognised only when words following them have themselves been identified (Bard *et al.*, 1988), words preceding a disfluent interruption may be hard to recognise. If an Original Utterance contains any words which have as their right context a disfluent interruption, rather than the word which would in fluent speech provide the key to their identification, such words might be more prone to remain unrecognised. On the other hand, many words prior to a disfluent interruption could still be recognised *immediately*, since the facilitatory effect on recognition of left context is present in both disfluent and fluent sentence onset cases. The words following a disfluent interruption might be expected to present different problems to the recognition process. The unexpected – and usually ungrammatical – change in the course of a disfluent utterance means that the words following a disfluency – the first words of the fluent continuation – might be expected to be harder to recognise immediately than words in a similar serial position in fluent utterances.

The word-level gating technique used in this experiment (see Chapter 2, Section 2.4) allows us to test on-line word recognition at the same time as eliciting judgements about oncoming disfluency.

In summary, a word-level gating experiment could test listeners' ability to detect cues to oncoming disfluency ("editing signals") in spontaneous English speech. The design of the experiment allowed other hypotheses to be examined, regarding the effect of the presence of disfluency on the recognition of words in its vicinity. Four main hypotheses were thus examined.

1. **Hypothesis 1:** listeners can detect oncoming disfluency when they hear an editing signal at the end of the original utterance, before the onset of the first word of the continuation.

2. **Hypothesis 2:** under the conditions of this experiment, listeners use non-recognition of words as cues to oncoming disfluency;

3. **Hypothesis 3:** if the last word of the original utterance, before a disfluent interruption, is not immediately recognised, it will be more likely to remain unrecognised than the word at a similar point in a fluent utterance.

4. **Hypothesis 4:** the first word of the continuation, directly after a disfluent interruption, will be less likely to be recognised immediately than a similar word in the same serial position in a fluent utterance.

4.1.1 Method

Materials and Design

Materials were 30 disfluent stimuli selected from a corpus of digitally recorded spontaneous speech. The selection and construction of test stimuli and controls is described in detail in Chapter 3, Section 3.3. The full set of experimental stimuli consisted of 120 utterances. Since the 120 utterances consisted of 60 spontaneous utterances and 60 rehearsed, the test items were divided into two complementary sets of 60, each containing 30 spontaneous and 30 rehearsed, to be presented to separate groups of subjects. Thus both groups of subjects would hear 15 spontaneous disfluent utterances, 15 spontaneous fluent utterances, 15 rehearsed fluent versions of the spontaneous disfluent utterances (which we will hereafter refer to as "rehearsed disfluent") and 15 rehearsed copies of the spontaneous fluent utterances and neither group would hear both the spontaneous version and the rehearsed version of the same item.

Each set of 60 items was blocked by speaker and recorded on a separate test tape.

Since 25% of the test items were the spontaneous disfluent utterances and the total number of test items per speaker was 20, the set of test items for any one speaker contained two disfluencies in one of the two test tapes and three in the other. Within the same test tape, the number of disfluent items per speaker alternated from 2 to 3 from speaker to speaker. The same applied to the other three sets of items (spontaneous fluent, rehearsed disfluent, rehearsed fluent).

Since there were six speakers in all, the total number of each test item type per speaker was 10.

In order to decide the order of presentation for the test items, five sets of four items were defined for each speaker. The head of each set, item "A", was the spontaneous disfluent test item; its rehearsed version, the imitation of the edited disfluent item was item "B"; the spontaneous fluent item (matched for structure, length and prosody with "A", as far as possible) was item "C" and its rehearsed version, item "D". For the same group of subjects, items A and B of a set were mutually exclusive because they had the same onset string, as were items C and D, which contained all the same words. For each set, one spontaneous and one rehearsed version were presented to the same group of subjects. So, for any set, for one group of subjects, either items A and D or B and C were presented. Since the disfluent and fluent members of a set were similar in structure, the presentation of members {A and D} and {B and C} of the same set were always separated by a minimum of two other items. Apart from these conditions, the order of presentation of items for the first group of subjects was random with respect to disfluency type and chronological order of occurrence in the original conversation. The order of presentation for the second group of subjects was the "mirror image" of that of the first in that, where a spontaneous member of a set of items (i.e. A or C) occurred in group one, its rehearsed version (B or D) occurred in group two and vice versa.

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 context. The test items were gated at the onset of each word as determined both auditorily and visually from the time-amplitude waveform. The test tapes were produced automatically by a computer programme, which avoided the problem of sound distortion at the end-points of each gated presentation by smoothly decreasing the intensity to zero over the last 1.5ms.

The presentation of a test item was preceded by the announcement of the item number, a tone and then the orienting section. Then the test item was presented incrementally, one word added at each presentation, each presentation being preceded by a tone and two seconds' silence and followed by 5.5 seconds'

silence.

Before the test items for each new speaker began, the new speaker was announced and about 10 seconds of speech by that speaker were presented in order to help familiarise the subjects with the voice.

The experiment was preceded on the tape by a thorough explanation with examples and a practice section consisting of three test items using material not included in the corpus.

Answer sheets were prepared for the experiment and for the introduction and practice session. For each test item a separate sheet was used which had printed on it the orienting section of speech for that item, a reminder of the scoring system the subjects were to use for their judgements about fluency, a grid for the answers to be written in and numbers to be circled as part of the test.

Subjects

Subjects were 20 members of the staff and student community of the University of Edinburgh. All were native speakers of English and could be expected to be familiar with the range of accents represented in the experimental materials. All reported having normal hearing.

Procedure

The subjects were divided into two groups of ten, each group hearing a different test tape.

The subjects were seated individually in listening booths and provided with the answer sheets and Revox 3100 semi-open stereo headphones. They were asked to listen carefully to each presentation of a stimulus and at the end of each gate to perform two tasks: first, they were to write the last word they had heard in the appropriate box on the answer sheet, using a new line in the grid for each presentation of the same test item; then they were to make a judgement, by circling a number on a scale of one to five, about whether they thought the utterance would continue fluently or not. They were asked to use the number "1" to indicate that they were sure the utterance would continue fluently, "5" to indicate that they were sure that the utterance would continue disfluently, "2"

or "4" to indicate a slight feeling either way and "3" to indicate that they did not know. They were encouraged to make a judgement about the fluency of the continuation even if they could not identify the latest word[1]. They were told not to alter their disfluency judgements after the tone warning of the next word had sounded and only to alter word judgements by rewriting the word in the current line below the original judgement. After the instructions had been played, the tape was stopped and the subjects asked if they had any queries. The practice session followed, after which the tape was stopped again and the answer sheets collected and checked to make sure that the subjects were following the instructions correctly. The subjects were once more asked if they had any queries. The experiment was then run in two sessions of approximately 45 minutes, separated by a break for refreshments.

At the end of the experiment, comments were invited from the subjects about how they had coped with the tasks.

Results I: disfluency judgements

Twenty subjects each gave disfluency judgements for each word in 60 stimuli of varying length, yielding a total of 9,570 judgements. The distribution of 1-5 disfluency judgements differed significantly between stimulus types ($\chi^2 = 107.77$, $df = 12$, $p < 0.0001$) with relatively fewer "fluent" judgements for the spontaneous disfluent stimuli than any of the controls and also significantly fewer "fluent" judgements for the spontaneous fluent stimuli than for the rehearsed stimuli ($\chi^2 = 32.65$, $df = 8$, $p = 0.0001$) (table 4.1). The distribution of disfluency judgements for the two rehearsed sets did not differ significantly.

The disfluency judgements of interest for the main analysis are those at the crucial point in the disfluent utterances – the word prior to the interruption – and the equivalent points in the control utterances. If it is true that disfluency is predictable from the characteristics of the speech signal prior to the onset of the first word of the continuation, disfluency judgements at this point in the disfluent stimuli will be significantly higher on the 1-5 scale than judgements for the fluent control points.

[1]Referred to as "disfluency judgements", henceforth

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Spontaneous Disfluent	871 *33.2%*	709 *27.1%*	605 *23.1%*	353 *13.5%*	82 *3.1%*	2620 *100%*
Spontaneous Fluent	905 *37.6%*	689 *28.6%*	469 *19.5%*	277 *11.5%*	70 *2.9%*	2410 *100%*
Rehearsed "Disfluent"	931 *42.9%*	599 *27.6%*	403 *18.6%*	205 *9.4%*	32 *1.5%*	2170 *100%*
Rehearsed Fluent	1027 *43.3%*	646 *27.3%*	429 *18.1%*	214 *9.0%*	54 *2.3%*	2370 *100%*
Marginal Totals	3734 *39.0%*	2643 *27.6%*	1906 *19.9%*	1049 *11.0%*	238 *2.5%*	9570 *100%*

Table 4.1. Experiment 1: disfluency judgement distribution by stimulus type: all words.

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Spontaneous Disfluent	81 *27.0%*	80 *26.7%*	53 *17.7%*	66 *22.0%*	20 *6.7%*	300 *100%*
Rehearsed "Disfluent"	145 *48.3%*	91 *30.3%*	41 *13.7%*	20 *6.7%*	3 *1.0%*	300 *100%*
Spontaneous Fluent	114 *38.0%*	95 *31.7%*	61 *20.3%*	25 *8.3%*	5 *1.7%*	300 *100%*
Rehearsed Fluent	141 *47.0%*	92 *30.7%*	42 *14.0%*	24 *8.0%*	1 *0.3%*	300 *100%*
Marginal Totals	481 *40.1%*	358 *29.8%*	197 *16.4%*	135 *11.3%*	29 *2.4%*	1200 *100%*

Table 4.2. Experiment 1: disfluency judgement distribution for last word of original utterance by stimulus type

The distribution of 1-5 disfluency judgements at the crucial point for the four stimulus sets differed significantly ($\chi^2 = 101.294$, $df = 12$, $p < 0.0001$), but there was no significant difference between the three control sets (table 4.2). For the disfluent stimuli there were more judgements of "4" or "5", indicating that subjects detected oncoming disfluency, than in the controls (28.7% of all judgements for the disfluent stimuli as opposed to between 7.7% and 10% for the controls), and fewer judgements of "1" (27% as opposed to between 38% and 48.3%). It should be noted at this point, though, that even though the difference between judgement distributions for the disfluent stimuli and the controls is significant, there are still only a small number of strong "disfluent" judgements (see figure 4.1).

A non-parametric analysis of variance (Friedman test) compared disfluency judgements at the crucial point in the four stimulus types for all subjects and all materials ($N = 300$). The differences in rank totals were found to be highly

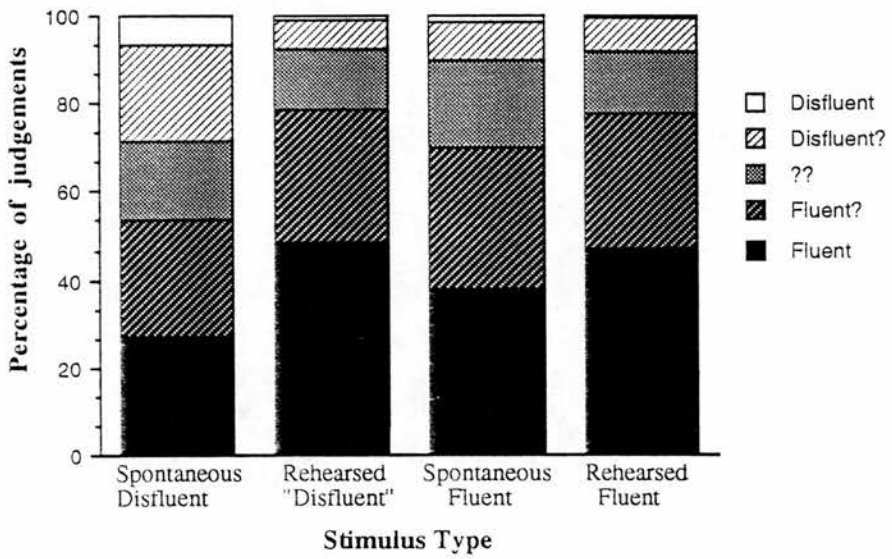

Figure 4.1. Experiment 1: fluency judgement distribution for last word of original utterance by stimulus type.

significant ($Xr^2 = 53.84$, $df = 3$, $p < 0.0001$), the highest rank total being for the spontaneous disfluent variable, showing that the disfluency judgements for disfluent stimuli at the crucial point were generally higher than those for the controls. Another Friedman test, omitting the spontaneous disfluent stimuli, also gave a significant result (at a lower level), showing that the spontaneous fluent stimuli also received higher judgements than the rehearsed controls ($Xr^2 = 7.56$, $df = 2$, $p = 0.023$). These results suggest two possible effects: the experimental hypothesis appears to be supported by a **fluency** effect – disfluent stimuli tended to be heard as more disfluent than the controls; and by a **mode** effect – spontaneous stimuli tended to be heard as more disfluent than rehearsed stimuli.

To examine these effects and any interactions more closely, parametric analyses of variance (ANOVAs) were used and the data were treated as interval data. Cells were made up of totals of judgements for each stimulus type, by subjects and by materials.

With these data, two-way ANOVAs with repeated measures for stimulus type were performed both by subjects and by materials, with **fluency** (disfluent vs fluent) and **mode** (spontaneous vs rehearsed) as independent variables. Significant main effects both by subjects and by materials were found for **fluency** ($F_{1(1,19)} = 14.62$, $p = 0.0011$; $F_{2(1,19)} = 4.84$, $p = 0.036$), for **mode** ($F_{1(1,29)} = 71.31$, $p < 0.0001$; $F_{2(1,29)} = 46.67$, $p < 0.0001$) and for the interaction of **fluency by mode** ($F_{1(1,19)} = 8.72$, $p = 0.0082$; $F_{2(1,29)} = 7.40$, $p = 0.0109$). $MinF'$ only reached significance for the mode effect (table 4.3). Figure 4.2 illustrates how the cell means differ for the four conditions: while the mean score for spontaneous disfluent stimuli is higher than that for the controls, it does not suggest that subjects were always convinced that disfluency was imminent.

Post hoc (Scheffé) tests showed that the fluency and mode main effects and the interaction were all caused by the higher disfluency scores for the spontaneous disfluent stimuli and that there was no significant effect of the difference between judgements for the spontaneous fluent and rehearsed fluent stimuli, nor of the difference between those for the two rehearsed sets of stimuli.

The results so far seem to support the experimental hypothesis to some extent: disfluent stimuli received significantly more "disfluent" judgements than

Source	F_1	df	α	F_2	df	α	$MinF'$	df	α
Fluency	14.62	1,19	0.0011	4.84	1,29	0.0359	3.63	1,44	ns
Mode	71.31	1,19	0.0001	46.67	1,29	0.0001	28.21	1,48	0.001
F×M	8.72	1,19	0.0082	7.40	1,29	0.0109	4.00	1,47	ns

Table 4.3. Experiment 1: F-ratios by subjects (F_1), by materials (F_2) and Minimum Quasi F-ratios ($MinF'$) for two-way ANOVAs with repeated measures for fluency and mode.

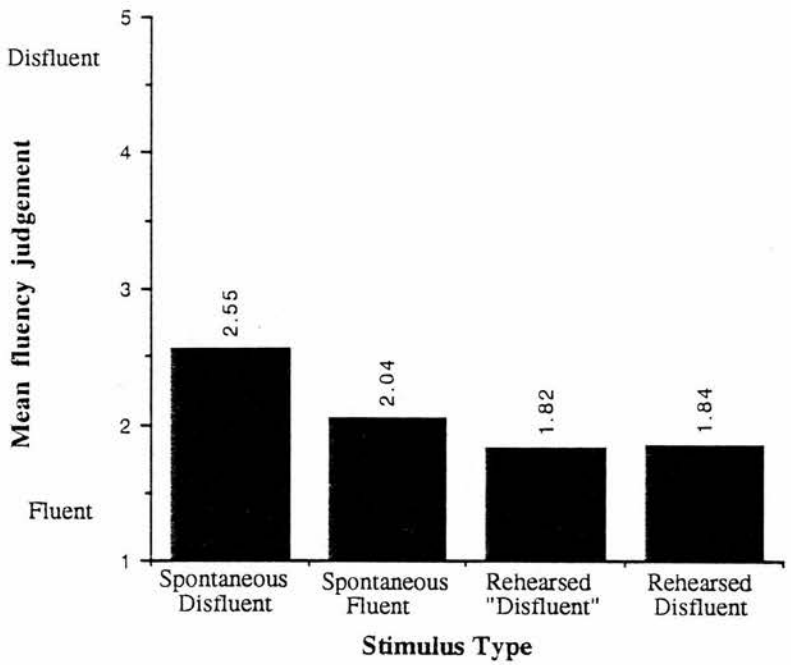

Figure 4.2. Experiment One: Means of fluency judgements at crucial word.

the controls and contribute most strongly to the fluency and mode effects observed in the analyses of variance. But it has been noted that the distribution of "disfluent" judgements does not give the impression that all subjects found all stimuli to contain cues to oncoming disfluency. In addition, there may be a mode effect, with spontaneous stimuli generally receiving fewer "fluent" judgements than rehearsed stimuli. Further analyses were required in order for these results to be understood more clearly: first, we needed to establish whether subjects were reacting to cues in the crucial word-gate or whether there were cues in the speech signal prior to this point; second, it would be of interest to know how subjects reacted to speech after the interruption point, where it might have been clear that disfluency had actually occurred, rather than that it was about to occur. For further investigation of the mode effect, it would also be useful to examine results at other gates in the presentation.

In this further set of analyses, the imminent disfluency judgements for the word prior to the interruption were compared with the judgements for the previous word and for the following word. If there were cues prior to the last word of the continuation, we would expect to see higher judgement totals for the previous word, as well as for the crucial word itself; if subjects adhered to their instructions and only gave higher judgements when they thought disfluency was *about* to occur, rather than when they thought it *had* occurred, then the judgements for the following word would be expected to be lower than for the crucial word. If, however, the mode effect was stable for fluent as well as disfluent stimuli, we would expect it to be apparent at points other than the crucial word.

Three-way ANOVAs with repeated measures were performed both by subjects and by materials with disfluency, mode and place (penultimate word of original utterance, final word of original utterance, first word of continuation) as independent variables. Significant effects both by subjects and by materials were found for **place** ($F_{1(2,38)} = 30.66,\ p < 0.0001$; $F_{2(2,58)} = 11.26,\ p = 0.0001$) and for **mode** ($F_{1(1,19)} = 100.53,\ p < 0.0001$; $F_{2(1,29)} = 28.60,\ p < 0.0001$), but not for **fluency**. The **place by mode** interaction was also significant ($F_{1(2,38)} = 30.61,\ p < 0.0001$; $F_{2(2,58)} = 19.78,\ p < 0.0001$). The **fluency by mode** interaction was only significant in the analysis by subjects ($F_{1(1,19)} = 9.85,\ p = 0.0054$) and the **place by fluency** and **place by fluency by mode** interactions did not

reach significance.

Means for the twelve variables in these ANOVAs are shown in table 4.4 and figure 4.3: it is unlikely that cues to oncoming disfluency were present *before* the last word of the interruption, since the mean value for the previous-word gate in the spontaneous disfluent stimulus set does not differ from the values for control stimuli; similarly, the mode effect (spontaneous, rehearsed) seems to be restricted to the crucial word, since means for the spontaneous fluent cell before and after this point do not differ from those of rehearsed stimuli. An additional and important observation to be made is that the mean of the disfluency judgement for the first word of the continuation in disfluent items is actually higher than that for the crucial word: since it is likely that rather than hearing cues to oncoming disfluency at this point, subjects were able to detect its actual *occurrence*, this finding casts some doubt over subjects' strategies in giving disfluency judgements, raising the question of whether they were responding to cues to oncoming disfluency, or to actual perception of occurring disfluency.

Post hoc Scheffé tests confirmed the first two of these observations, showing that all effects and interactions that reached significance were caused by the higher level of judgements for just two cells: judgements for the spontaneous disfluent stimuli at and after the crucial word. There were found to be no significant differences between any stimulus type pairs for the word before the last word of the original utterance. The mean of judgements in disfluent stimuli for the word after the interruption was higher than that for the crucial word (the word at the end of the original utterance), but this difference was not great enough to affect the place effect.

While the results of the analyses of variance seem to support the experimental hypothesis, other factors make the acceptance of this hypothesis less attractive. As we have seen, the mean of disfluency judgements for the first word of the continuation, at which point disfluency *has* occurred, is higher than the mean of judgements for the crucial word itself. In addition to this, the distribution table (table 4.2) shows that only 28.7% of judgements at the crucial word in disfluent stimuli were indications of the detection of oncoming disfluency while 53.7% of judgements indicate that subjects thought that the utterance would continue fluently. This is reflected in the means for the spontaneous disfluent

Variable	$\bar{X}$	sd_s	sd_m
SD1	1.95	0.470	0.638
SD2	2.55	0.509	0.680
SD3	2.67	0.532	0.543
SF1	1.81	0.420	0.576
SF2	2.04	0.479	0.633
SF3	1.94	0.367	0.715
RD1	1.79	0.455	0.390
RD2	1.82	0.431	0.582
RD3	1.96	0.400	0.692
RF1	2.05	0.506	0.616
RF2	1.84	0.461	0.396
RF3	1.81	0.437	0.424

Table 4.4. Experiment 1: Cell means and standard deviations for 3-way ANOVAs (Place by fluency by mode). S = "Spontaneous", R = "Rehearsed", D = "Disfluent", F = "Fluent" and 1,2,3 are places, before, at and after the crucial word.

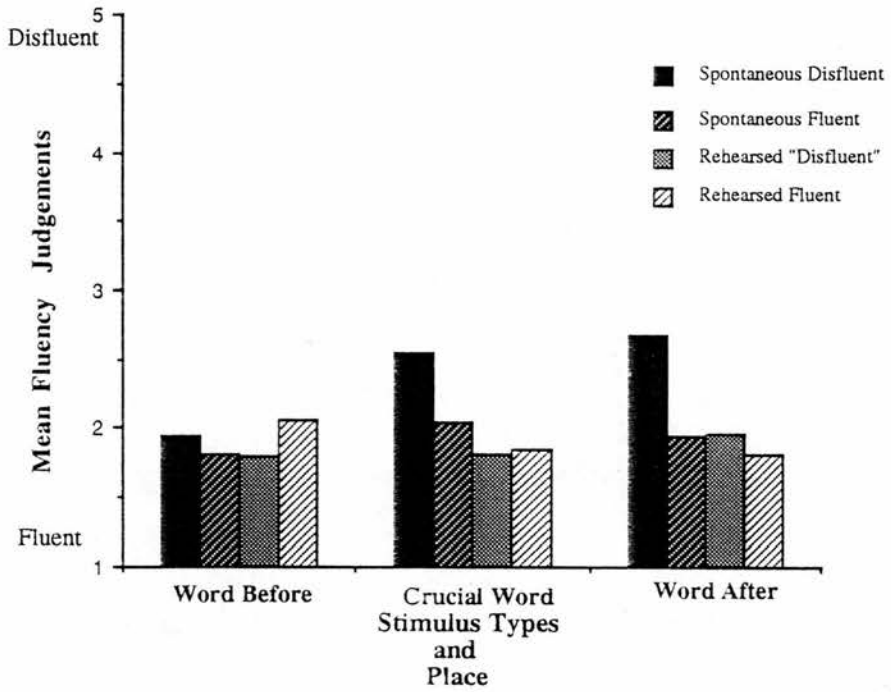

Figure 4.3. Experiment 1: Cell means for 3-way ANOVAs (Place by fluency by mode)

cell judgement (table 4.4), which is lower than would be expected for a positive identification of oncoming disfluency, while the level of means for the controls *are* at the expected level for fluent continuations. Given that subjects reacted to disfluency that had already occurred by giving higher disfluency judgements, it may be that the higher means for the crucial word were subjects' responses to what they actually perceived as disfluency in the speech signal, rather than some discrete editing signal. Possible such percepts are words that are clearly broken off before their ending (fragments) and words followed by perceivable silent pauses.

If subjects *did* react to fragments and silent pauses as cues to disfluency, then dividing the data into two separate sets, those for stimuli containing such cues and those for stimuli without them, should create an interaction with fluency for spontaneous examples. Twelve stimuli contained a silent pause at the interruption point of over 100ms duration; a partially overlapping set of 8 stimuli had original utterances ending in a fragment. The total number of stimuli containing either or both of these cues was 18 and the number without cues, 12.

Ths distribution of imminent disfluency judgements at the crucial gate in spontaneous disfluent stimuli differed significantly between the with-cue and the no-cue condition, with more "fluent" and fewer "disfluent" judgements in the no-cue condition ($\chi^2 = 30.86$, $df = 4$, $p < 0.0001$) (table 4.5). The distribution of judgements in the no-cue condition did not differ from that for the control stimuli (figure 4.4)

Analyses of variance were performed by subjects, as a three-way ANOVA with repeated measures for the binary conditions of fluency, mode and cue, and by materials, as a two-way ANOVA with repeated measures for fluency and mode, with cue as a grouping factor. Cell means (by subject) are illustrated in figure 4.5. Significant main effect of the presence of **cue** was found by subjects and by materials ($F_{1(1,19)} = 16.06$, $p < 0.001$; $F_{2(1,28)} = 7.29$, $p = 0.0116$; $MinF'_{(1,45)} = 5.01$, $p < 0.05$); a strong **mode** effect was found in both analyses, too ($F_{1(1,19)} = 63.46$, $p < 0.0001$; $F_{2(1,28)} = 42.65$, $p < 0.0001$; $MinF'_{(1,46)} = 25.51$, $p < 0.01$), but the **fluency** effect only reached significance in the by-subjects analysis ($F_{1(1,19)} = 10.25$, $p = 0.0047$). The **fluency by mode** interaction reached significance in both analyses ($F_{1(1,19)} = 8.20$, $p < 0.01$;

Cue ?	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Absent	46	40	18	14	2	120
%	*(38.3%)*	*(33.3%)*	*(15.0%)*	*(11.7%)*	*(1.7%)*	*(100%)*
Present	35	40	35	52	18	180
%	*(19.4%)*	*(22.2%)*	*(19.4%)*	*(28.9%)*	*(10.0%)*	*(100%)*
Marginal Totals	81	80	53	66	20	300
	(27.0%)	*(26.7%)*	*(17.7%)*	*(22.0%)*	*(6.7%)*	*(100%)*

Table 4.5. Experiment 1: fluency judgement distribution by presence of cue in spontaneous disfluent stimuli.

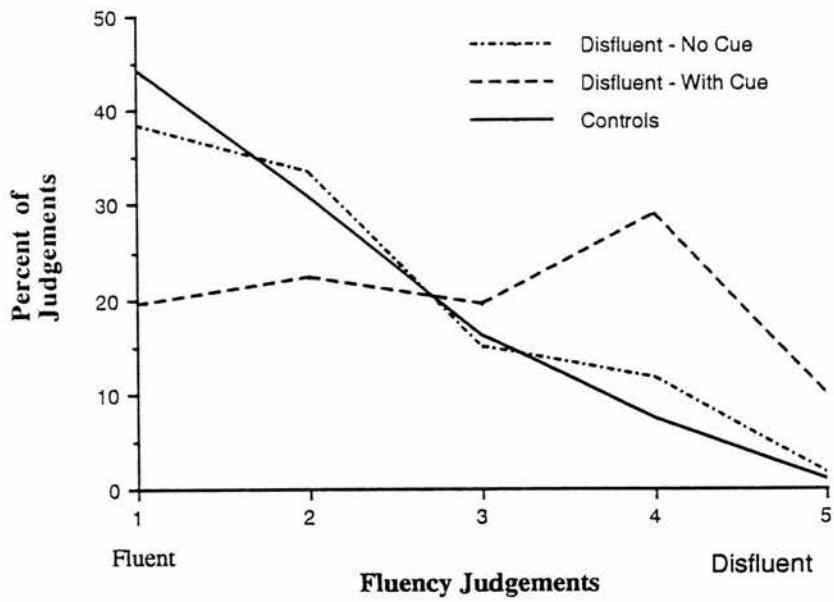

Figure 4.4. Experiment One: Percentage of judgements for crucial word in presence and absence of cue (pause or fragment), compared to controls.

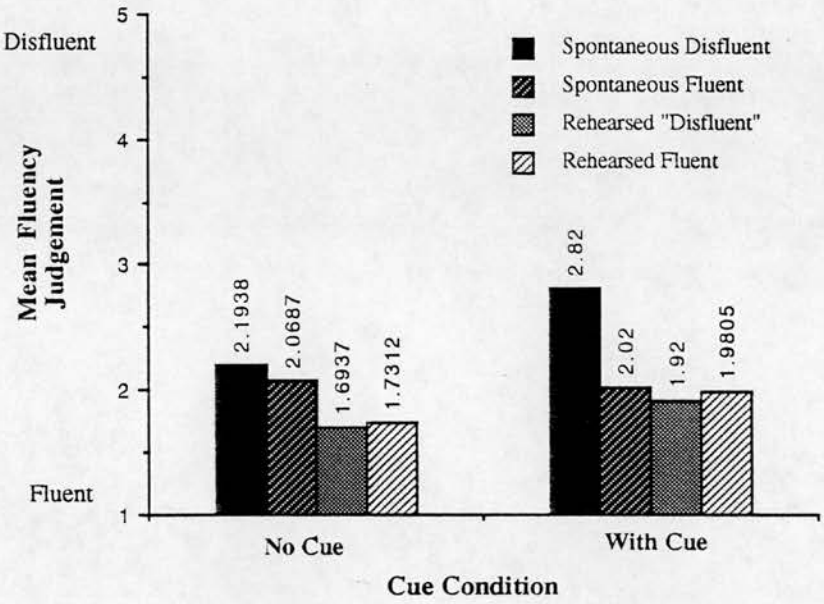

Figure 4.5. Experiment One: Effect of cues on fluency judgements.

by cue interactions reached significance in the by-subjects analyses only (F×C: $F_{1(1,19)} = 10.34$, $p = 0.0046$; F×M×C: $F_{1(1,19)} = 7.15$, $p = 0.015$).

Scheffé tests verified the impression given by figure refex1-fig.fluj.fmc: the spontaneous disfluent with-cue mean significantly exceeds all other means including, most importantly, the spontaneous disfluent no-cue cell. All means for spontaneous cells were higher than all means for rehearsed cells, but Scheffé tests suggested that the mode effect was mainly due to differences between the spontaneous disfluent cell and all other cells. The **fluency** effect found in the by-subjects analysis was also found to be due to the higher judgements for the spontaneous disfluent cell in the with-cue condition: the level of the spontaneous disfluent no-cue mean was not significantly higher than its fluent counterpart ($p > 0.05$).

So the presence of cues like a silent pause or a word fragment in the crucial gate led to significantly higher mean fluency judgements than were found in disfluencies with no such cues. Where such cues were absent, the mean fluency judgements did not differ significantly from their spontaneous fluent controls.

judgements did not differ significantly from their spontaneous fluent controls.

Finally, disfluency judgements for the crucial point in each disfluent stimulus were compared with those for the matched spontaneous fluent control in Wilcoxon signed rank tests. It was found that judgements for the disfluent condition were significantly ($p < 0.05$) higher than those for the fluent condition in only 12 of the 30 cases, the difference in scores was insignificant in 15 cases and the difference was significantly higher for the fluent condition in 3 cases. Of the set of 18 disfluent stimuli defined as containing a cue within the crucial gate, 9 had significantly higher disfluency judgements than their matched fluent pairs, 8 did not differ significantly and in one case the fluent stimulus yielded higher disfluency judgements; of the 12 disfluent stimuli with no cue, 3 yielded higher judgements than their pairs, 8 did not differ significantly and 2 had lower judgements.

Results II: Word recognition

The analysis central to Experiment One treated subjects' ability to detect oncoming disfluency. But subjects also had to perform a word recognition task. At each presentation, they were asked to write down the last word they had heard and to make any corrections necessary to previous words in the test utterance using the appropriate boxes on the answer sheet.

The results of the word-recognition task are of interest from two points of view: there may be a correlation between judgements of oncoming disfluency and performance in the word-recognition task, suggesting that when subjects have trouble recognising a word they are less sure about the fluency of the utterance; on the other hand, a weaker performance around a point of disfluency may indicate an effect of the presence of disfluency on word-recognition and thus on utterance processing as a whole.

The results of performance in the word-recognition task and related results are presented in three stages. First, the overall word-recognition performance and the differences in word-recognition performance among the four different types of utterance are examined. Next, the word-recognition performance in the vicinity of disfluency is compared with the disfluency judgements at these points to examine the effect of failed word-recognition on disfluency judgements: here

we examine the hypothesis that subjects gave more "disfluent" judgements when they were unable to recognise the current word (Hypothesis 2). Finally, the effect of the presence of disfluency on subjects' ability to recognise words is examined by comparing word-recognition performance in the vicinity of disfluency with the performance at the equivalent points in the fluent controls: under Hypothesis 3, if the word before a disfluent interruption is not immediately recognised, it will be more likely to remain unrecognised than similar words in fluent stimuli; under Hypothesis 4, the word immediately following the interruption will be less likely to be recognised on first presentation than a similar word in the same serial position in a fluent utterance.

Overall word recognition performance. The 120 utterances presented in each experiment contained a total of 957 word tokens. Each word token was presented to 10 subjects, giving a total of 9570 recognition outcomes for each of the two experiments. Each recognition outcome was classified as "immediate", where the word was recognised correctly on its first presentation, "late", where the word was recognised on a subsequent presentation, or "missed", where the word remained unrecognised. In assessing subjects' responses, the original word and its homophones, with or without correct inflection, were scored as correct recognitions.

Overall, 8048 (84.1%) of the 9570 recognition outcomes were immediate, 1004 (10.5%) were late and 518 (5.4%) were missed.

For the spontaneous disfluent stimuli, the total number of outcomes was 2620, of which 2138 (81.6%) were immediate recognitions, 266 (10.2%) late and 216 (8.2%) missed. In the spontaneous fluent utterances, the total number of outcomes was 2410. Of these, 2042 (84.7%) were immediate recognitions, 251 (10.4%) were late and 117 (4.9%) were missed. For the rehearsed versions of the disfluent stimuli, there were a total of 2170 outcomes, of which 1804 (83.1%) were immediate recognitions, 278 (12.8%) were late and 88 (4.1%) missed (table 4.6). For the rehearsed fluent stimuli, the total number of outcomes was 2370, of which 2064 (87.1%) were immediate recognitions, 233 (9.8%) late and 73 (3.1%) missed. The distribution of word recognitions differed significantly between stimulus types, the greatest difference being in higher miss-rates in spontaneous disfluent stimuli

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	2138	2042	1804	2064	8048
%	*81.6%*	*84.7%*	*83.1%*	*87.1%*	*84.1%*
Late	266	251	278	233	1028
%	*10.2%*	*10.4%*	*12.8%*	*9.8%*	*10.7%*
Missed	216	117	88	73	494
%	*8.2%*	*4.9%*	*4.1%*	*3.1%*	*5.2%*
Total	2620	2410	2170	2370	9570

Table 4.6. Experiment 1: Word recognition outcomes for all words in all stimuli.

than in the controls ($\chi^2 = 90.088$, $df = 6$, $p < 0.0001$). Within the controls, the distribution of outcomes for rehearsed "disfluent" stimuli differed significantly from those for the other sets ($\chi^2 = 21.745$, $df = 4$, $p = 0.0002$): the rehearsed "disfluent" set included more late recognitions. The spontaneous and rehearsed fluent stimuli also differed significantly in the distribution of word recognition outcomes, with more missed recognitions in the spontaneous set than in the rehearsed set ($\chi^2 = 10.643$, $df = 2$, $p = 0.0049$).

Effect of failed word recognition on disfluency judgements. Given that the disfluent stimuli yielded more "missed" word recognition outcomes, it is possible that the greater number of judgements of "disfluent" here mark subjects' reactions to their inability to recognise a word.

To test this hypothesis, disfluency judgements for words recognised at first presentation were compared with those for other words. Comparisons were made for the crucial word and for the word following the interruption, since these two words are the words which received most "disfluent" judgements. Both sets of spontaneous stimuli were tested.

The distribution of disfluency judgements was not significantly associated with word recognition outcomes for either set of stimuli or for either word (before

or after interruption point) ($p > 0.06$). The null hypothesis, that there is no difference between disfluency judgements for words recognised immediately and words not recognised immediately, is therefore not rejected.

Effect of disfluency on word recognition. Little is known about the effect of the presence of disfluency on listeners' ability to process spontaneous speech. This experiment made it possible to examine one possible manifestation of the effect, namely the effect on the ability of listeners to recognise words in the vicinity of a disfluent interruption. As above, we concentrate on the words immediately prior to and immediately following the disfluent interruption: comparisons are made with word recognition outcomes at the equivalent points in all controls. As we noted in section 4.1, the presence of disfluency may affect the recognition of either of these two words for different reasons. In the case of the word prior to the interruption, the full left context is present and the right context missing: since the right context is often needed for successful word recognition to be achieved (Bard *et al.*, 1988), we might expect more missed recognitions for the word prior to disfluent interruption. In the case of the word following the interruption, there is no coherent immediate left context but, in most cases a coherent right context: this may result in more late recognitions, as is typical for words nearer the beginning of utterances; in addition, the fact that the utterance is unexpectedly interrupted may hamper immediate word recognition.

For the word prior to the interruption, the distribution of word recognition outcomes differed significantly with stimulus type ($\chi^2 = 78.76$, $df = 6$, $p < 0.0001$). The word was recognised immediately in the spontaneous disfluent stimuli in 79.3% of cases ($N = 300$) and missed in 14.7%, as opposed to averages of 91.9% immediate recognitions and 2% missed recognitions for the matched point in all controls. The distribution of late recognitions did not differ significantly between spontaneous disfluent stimuli and the controls (table 4.7). However, in 8 of the 30 disfluent stimuli, the original utterance ended in an incomplete word. These words had been treated in the same manner as full words in the analysis and marked as "recognised" when correctly recognised as fragments of words: of a total of 80 recognition outcomes, 34 (42.5%) were recognised on first presentation, 14 (17.5%) were recognised late and 44 (40%) were missed. To assess the

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	238	282	266	279	1065
%	*79.3%*	*94.0%*	*88.7%*	*93.0%*	*88.8%*
Late	18	15	24	16	73
%	*6.0%*	*5.0%*	*8.0%*	*5.3%*	*6.1%*
Missed	44	3	10	5	62
%	*14.7%*	*1.0%*	*3.3%*	*1.7%*	*5.2%*
Total	300	300	300	300	1200

Table 4.7. Experiment 1: Word recognition outcomes for word before disfluent interruption in all stimuli.

effect of the presence of disfluency on full words at the crucial point, the judgements for fragments were removed from the analysis. The resulting distribution of word recognition outcomes still differed significantly between stimulus types, but at a lower level of significance and with a different pattern of outcomes in the spontaneous disfluent case ($\chi^2 = 13.22$, $df = 6$, $p = 0.0396$) (table 4.8). The stimulus sets did not differ with respect to the number of immediate recognitions when fragments were excluded from the disfluent set, but the number of missed recognitions was significantly greater in the disfluent cases, and there were more "late" recognitions in the controls. This result supports Hypothesis 3 (page 74): if the last word of the original utterance, before a disfluent interruption, is not immediately recognised, it will be more likely to remain unrecognised than the word at a similar point in a fluent utterance.

For the word following the interruption, there was also a significant difference in the distribution of word recognition outcomes between the stimulus sets ($\chi^2 = 54.23$, $df = 6$, $p < 0.0001$). The disfluent stimuli yielded fewer immediate recognitions than the controls (77.7% versus an average of 91.67% for all controls) and more late and missed recognition outcomes (13.0% late and 9.3% missed versus 5.57% and 2.8%) (table 4.9). The presence of a preceding fragment made

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	199	204	199	201	807
%	*92.7%*	*92.3%*	*90.5%*	*91.4%*	*91.7%*
Late	4	14	14	14	46
%	*1.8%*	*6.4%*	*6.4%*	*6.4%*	*5.2%*
Missed	12	3	7	5	27
%	*5.5%*	*1.4%*	*3.2%*	*2.3%*	*3.1%*
Total	220	220	220	220	800

Table 4.8. Experiment 1: Word recognition outcomes for word prior to interruption in all stimuli with complete words.

no difference to the distribution of word recognition outcomes for this word. The spontaneous fluent stimuli yielded a distribution of word recognition outcomes that differed significantly from the other controls in this analysis, with a higher rate of immediate recognitions and fewer late recognitions than the rehearsed stimuli ($\chi^2 = 13.78$, $df = 4$, $p = 0.008$).

Since the word after the interruption does not have a coherent immediate left context it might be viewed as being similar in potential for immediate recognition to utterance-initial words, which are known to be less likely to be recognised immediately than words later in the utterance (Bard *et al.*, 1988; Pickett & Pollack, 1963; Pollack & Pickett, 1963; Pollack & Pickett, 1964). For this reason, it was interesting to examine the word recognition outcomes at the onset of the stimuli and then to compare them with the distribution of word recognition outcomes at the first word in the continuation.

Initially, to confirm that the serial position effect on word recognition applied to the present set of data, the percentage of immediate recognitions at each of gates 1-11 in all three sets of control stimuli was compared with its serial position: the serial position effect was confirmed ($r = 0.794$, $df = 32$, $p < 0.0001$) and the overall mean percentage of immediate recognitions for the stimulus-initial word

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	233	267	286	272	1058
%	*77.7%*	*89.0%*	*95.3%*	*90.7%*	*88.2%*
Late	39	25	5	20	89
%	*13.0%*	*8.3%*	*1.7%*	*6.7%*	*7.4%*
Missed	28	8	9	8	53
%	*9.3%*	*2.7%*	*3.0%*	*2.7%*	*4.4%*
Total	300	300	300	300	1200

Table 4.9. Experiment 1: Word recognition outcomes for word after interruption in all stimuli.

was the lowest.

Next, a comparison was made of the distribution of recognition outcomes for the first word in the presentation of each of the four types of stimuli. The distribution was found to differ significantly between the four sets of stimuli ($\chi^2 = 56.41$, $df = 6$, $p < 0.0001$). There was no significant difference between word recognition outcomes for the spontaneous and rehearsed fluent sets (consisting of the same set of words), but results for the rehearsed "disfluent" set differed significantly from the other control sets ($\chi^2 = 45.76$, $df = 4$, $p < 0.0001$), with fewer immediate and more late recognitions. The spontaneous disfluent and rehearsed "disfluent" sets also differed significantly ($\chi^2 = 11.79$, $df = 2$, $p = 0.0028$); the difference was not in the number of immediate recognitions but in the distribution of outcomes between late and missed, with more missed words and fewer late recognitions in the spontaneous cases. It is likely that this difference in distribution of outcomes is a consequence of the presence of disfluency near the beginning of the onset of some of the spontaneous disfluent stimuli, which, as we have seen above, can inhibit the late recognition of words which are not recognised immediately.

Finally, the distribution of word recognition outcomes for the word following

the interruption in spontaneous disfluent stimuli was compared to the distribution of utterance-initial outcomes. The distribution of outcomes for the word after the interruption happened to be identical, in numbers of immediate, late and missed recognition, to that of the outcomes for the first word in spontaneous fluent stimuli. We can conclude from this that the word after a disfluent interruption is likely to present similar problems to the process of word recognition as the first word in a new utterance: it is more likely to be recognised later than words at the same serial position in a fluent utterance because, like utterance initial words, it lacks the left sentential context that appears to aid the immediate recognition of words later in a fluent utterance. This result supports Hypothesis Four.

4.1.2 Discussion

The experiment presented a sample of spontaneous disfluent utterances and three sets of controls in word-level gating format, in order to address three main questions regarding the understanding of disfluent speech. The first question was inspired by work in computational linguistics (Hindle, 1983) which relies on the detection of an editing signal in the speech stream for successful recognition of disfluency. The other questions concerned listeners' ability to recognise words in the immediate vicinity of a disfluent interruption and the relationship between word and disfluency recognition.

Detecting oncoming disfluency

The first hypothesis, and the most important from the point of view of this thesis, was that listeners would be able to detect oncoming disfluency before the onset of the first word of the continuation. If the hypothesis was supported it would provide psycholinguistic evidence for the notion that listeners make use of a discrete editing signal placed at the end of the original utterance.

Disfluent stimuli overall and particularly at the crucial word yielded fewer "fluent" and more "disfluent" judgements than the controls. The analyses of variance showed that the means of disfluency judgements for the crucial word were significantly higher than for equivalent points in all the controls.

These results appeared to support the experimental hypothesis, but other

analyses argue for a different view of the outcome. First, it was found that disfluency judgements for the first word of the continuation were higher than for the crucial word itself: this suggested that subjects might have been giving higher (more "disfluent") judgements on the basis of having already noticed the presence of a disfluency, rather than having detected a signal of oncoming disfluency. Second, two features of a subset of the disfluent stimuli – the presence of silent pause at the interruption point and incomplete words – were identified as being possible cues which could either be interpreted as constituting disfluency themselves or as being clear cues to oncoming discontinuity. In fact, predictions of disfluency exceeded controls *only* where such cues were present (18 sentences); without these cues, neither the distribution of disfluency judgements nor the mean judgement differed significantly from those of the spontaneous fluent controls.

For individual stimuli, only a minority (12) of the 30 disfluent stimuli differed from their spontaneous fluent controls in a way which directly supported Hypothesis One and 3 cases actually gave significant results in direct opposition.

So, while the initial analyses seemed to suggest that the experimental hypothesis was supported, subsequent examination, which took into account the two factors word-fragmentation and silent pause, showed that in the absence of such cues, subjects gave lower disfluency judgements, which did not differ from those given for spontaneous fluent controls, suggesting that they were unable to predict the oncoming disfluency. Responses for the first word of the continuation showed that subjects often judged stimuli to be "about to be disfluent" when in fact disfluency had already occurred and the continuation had commenced. It is possible that the fragments and silent pauses could themselves be seen as constituting disfluency in that the expected continuation (word-completion or continuation of speech within the expected time frame) could already be heard not to be present in the gated presentations which contained them, and thus that the greater incidence of "disfluent" judgements for such cases was for the same reasons as their greater incidence for the first word of the continuation – subjects were reacting to perceived disfluency, rather than to some editing signal which alerted them to an oncoming disfluency. Thus we are unable to reject the null hypothesis.

Another finding of interest to the perception of disfluent speech is that the

word prior to the last word of the original utterance was not found to contain any perceptually useful cue to oncoming disfluency: the only such cues apparent from the results were in the gates in the immediate vicinity of the interruption.

The main finding of this experiment, that subjects cannot reliably detect an editing signal at the end of the original utterance in disfluent speech, has obvious consequences for a model of speech understanding which relies on the detection of such a signal (Hindle, 1983). Apart from this study, no other published studies known to the author have attempted to find psycholinguistic evidence to support the editing signal hypothesis. No acoustic-phonetic correlates of the hypothesised discrete editing signal have been established and Hindle's use of Labov's notion of *"an abrupt cutoff in the speech signal"* seems to be most appropriate in the case of fragments, but hard to apply elsewhere. Other more recent studies in computational linguistics (Bear *et al.*, 1992; Shriberg *et al.*, 1992; Nakatani & Hirschberg, 1993a; Nakatani & Hirschberg, 1993b) (studies which postdate the present experiment and the publication of its results in (Lickley *et al.*, 1991)) have attempted to address the issue in the light of the elusiveness of the editing signal. These were reported in Chapter 2.

Word recognition and disfluency

The second task that subjects had to perform in the experiment was word recognition. Word recognition outcomes for each presentation to each subject were classed as either "immediate", "late" or "missed".

The task provided data to control for a possible artefact in the disfluency detection task and to test the effect of disfluency on subjects' ability to recognise words.

It was possible that the higher rates of judgements of "disfluent" found for the spontaneous disfluent stimuli were an artefact of subjects' failure to recognise words in the vicinity of a disfluent interruption. It was found that the two words on either side of the interruption had less chance of being recognised than similar words in the fluent controls, but no significant relationship was found between late or missed recognition and judgements of "disfluent".

Two hypotheses regarding word recognition were examined in the experiment.

First, it was predicted that the word at the end of the original utterance, directly before the interruption would be equally likely to be recognised on first presentation as similar words in similar positions in fluent stimuli, but that if a word was not recognised immediately, it would be less likely to be recognised given following context than a word with a fluent right context. Second, it was predicted that the word following a disfluent interruption would be more likely to require right context for its successful recognition than a word which was in the same serial position in a fluent utterance.

Both of these hypotheses were supported by the results. For the word before the interruption, where that word was not a fragment, the rate of immediate recognitions did not differ from that for fluent controls but the rate of "missed" recognitions was significantly higher. For the word after the interruption, the number of immediate recognitions was significantly less than for similar words in the controls and the distribution of recognition outcomes did not differ from that found for words at the beginning of the stimuli.

These results add support to the findings of Bard, Shillcock and Altmann, 1988 (Bard *et al.*, 1988), that words are often recognised after their acoustic offset in spontaneous speech. The results in this experiment do not match very closely those of the other researchers, who found that of all successful recognitions 21% occurred after the end of the word in question: our results show that 11% of successful recognitions for the pooled spontaneous stimuli were late. This difference may be due to a combination of factors: the recording conditions for the data used in this experiment were very carefully controlled, resulting in digital recordings of very high quality; the sampling rate of 20KHz for the materials maintained this high standard; the speakers who provided our corpus were chosen as being speakers of "close to standard" British English. It may be, therefore, that the materials used in this experiment were closer to "lab speech" than those used in (Bard *et al.*, 1988), which did not have such idealised recording conditions, used a lower sampling rate and came from a larger number of less "standard" speakers and, as a result, probably resembles real-life listening conditions more than the present study.

From the point of view of computational models, the word recognition results add to the barrage of results which dispute the psychological reality and practical

application for on-line processing of CL models which assume word recognition can be performed by bottom up information alone in spontaneous speech. The finding that words in the immediate vicinity of a disfluency are harder to recognise than similar words in fluent speech and particularly the finding that the word before an interruption is more likely not to be recognised at all, add to the difficulties for a model which relies primarily on syntactic cues to detect and filter out disfluencies: if the words aren't all recognised, syntactic anomalies will be hard to identify correctly.

4.2 Experiment 2: Detecting Disfluency

Experiment One showed that prior to the interruption point in disfluent utterances listeners do not have access to information which warns them of oncoming disfluency. Subjects reacted to cues such as incomplete words and long pauses with judgements of "disfluent" where such cues were present in the word-gate prior to the onset of the fluent continuation, but gave even more "disfluent" judgements for the word which constituted the beginning of the continuation, whether or not the utterance contained a cue. This suggested that their judgements were more reactions to *perceived* disfluency itself, rather than to an editing signal which predicted oncoming disfluency. If so, subjects were often able to detect disfluency before the end of the first word of the continuation. In order to test this hypothesis, an experiment was designed to find out how soon listeners could detect the *presence* of disfluency when explicitly instructed to do so.

No previous psycholinguistic studies have produced empirical evidence regarding the question of how soon in the processing of a disfluent utterance listeners are able to detect that disfluency has occurred. Levelt (1983) suggests that a combination of syntactic and lexical constraints and cues such as editing terms and discontinuous sentence prosody make it theoretically possible for listeners to detect disfluency promptly and solve the problem of connecting the repair to the original utterance.

Computational models, basing their approach to disfluency detection mainly on syntactic information, would often need more than just a single word before being able to detect and resolve the parsing problems caused by disfluency (discussed in Chapter 2, Section 2.2.2)

Levelt's observations that the detection of disfluency is in principal possible within the first word of the repair, together with evidence from Experiment One that listeners were able to detect disfluency at this point, form the basis of the main hypothesis for Experiment Two. The same materials and the same word-gating procedure as used in Experiment One were used for this experiment. Subjects again performed dual tasks, but in this case disfluency detection was to be

done in tandem with word recognition. It was thus possible to test the word-recognition hypotheses examined in the first experiment on a second group of subjects, testing the relationship between failure to recognise words and disfluency judgements.

In summary, a word-level gating experiment tested listeners' ability to detect disfluency in a sample of utterances from a corpus of spontaneous English speech. The experimental procedure also allowed us to test for the relationship, if any, between word recognition and judged disfluency. Four main hypotheses were tested, the last three having also been tested in Experiment One.

1. **Hypothesis 1:** listeners can detect disfluency by the offset of the first word of the continuation, which immediately follows the interruption;

2. **Hypothesis 2:** under the conditions of this experiment, listeners react to failure to recognise a word by judging the stimulus to be disfluent;

3. **Hypothesis 3:** if the last word of the original utterance, before a disfluent interruption is not immediately recognised, it will be more likely to remain unrecognised than the word at a similar point in a fluent utterance;

4. **Hypothesis 4:** the first word of the continuation, directly after a disfluent interruption, will be less likely to be recognised immediately than a similar word in the same serial position in a fluent utterance.

4.2.1 Method

Materials and design

The materials used in this experiment were identical to those used in the first experiment. The introduction and the practice items were changed to give instructions and adequate practice on the new task and the answer sheet was altered slightly to reflect the difference in the disfluency judgement task. The design was thus also identical to that of Experiment One.

Subjects

Subjects were 20 members of the staff and student community of the University of Edinburgh. All were native speakers of English and could be expected to be familiar with the range of accents represented in the experimental materials. All reported having normal hearing.

Procedure

The procedure for this experiment was mainly the same as for that of the first experiment, with the same number of subjects split into two groups of ten, each group hearing a different test tape, and the experiment being run in two sessions of approximately 45 minutes.

The subjects were asked to perform the word recognition task in the same manner as in the first experiment. The disfluency judgement task differed: subjects were asked to make a judgement on a one to five scale about whether they considered that the utterance was fluent at the current word gate. The number 1 was used to indicate that the subject considered that the utterance was fluent at that point, the number 5 to indicate that they considered that the utterance was disfluent, numbers 2 and 4 to indicate a slight feeling either way and number 3 to indicate that they did not know. The subjects were asked to make their judgement about fluency at a given word with respect to that word's relationship with the previous word only: so if, following a disfluency, the utterance continued fluently, then the judgements in the continuation should not be affected by the earlier disfluency. In cases where subjects wished to alter their disfluency judgement at a given point having heard a subsequent word or words, they were asked to indicate this by drawing an asterisk in the column corresponding to the original judgement but on the current line of the answer grid: it was stressed that this was particularly important where a disfluency had been recognised later than at the first word of the continuation.

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Spontaneous Disfluent	1640 *62.6%*	230 *8.8%*	209 *8.0%*	169 *6.5%*	372 *14.2%*	2620 *100%*
Spontaneous Fluent	1861 *77.2%*	179 *7.4%*	138 *5.7%*	105 *4.4%*	127 *5.3%*	2410 *100%*
Rehearsed "Disfluent"	1725 *79.5%*	145 *6.7%*	137 *6.3%*	73 *3.4%*	90 *4.1%*	2170 *100%*
Rehearsed Fluent	1858 *78.4%*	199 *8.4%*	129 *5.4%*	82 *3.5%*	102 *4.3%*	2370 *100%*
Marginal Totals	7084 *74.0%*	753 *7.9%*	613 *6.4%*	429 *4.5%*	691 *7.2%*	9570 *100%*

Table 4.10. Experiment 2: disfluency judgement distribution by stimulus type: all words.

4.2.2 Results I: disfluency judgements

Twenty subjects gave disfluency judgements on each word in 60 stimuli of varying lengths, yielding a total of 9,570 judgements. The distribution of 1-5 disfluency judgements differed significantly between stimulus types ($\chi^2 = 367.808$, $df = 12$, $p < 0.0001$), with the greatest differences being between the spontaneous disfluent stimuli and all controls for the percentage of judgements of "fluent" ("1") (62.6% in the disfluent stimuli versus an average of 78.4% for the controls) and of judgements of "disfluent" ("5") (14.2% versus 4.6%). The distribution of judgements did not differ significantly between the three control conditions (table 4.10).

In this experiment, the disfluency judgements of greatest interest are those for the word which constitutes the interruption or directly follows the interruption point. If disfluency is recognised at this point, we expect to see most subjects giving judgements of "4" or "5" for disfluent stimuli and lower judgements for

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Spontaneous Disfluent	42 *14.0%*	13 *4.3%*	22 *7.3%*	37 *12.3%*	186 *62.0%*	300 *100%*
Spontaneous Fluent	272 *90.7%*	12 *4.0%*	4 *1.3%*	5 *1.7%*	7 *2.3%*	300 *100%*
Rehearsed "Disfluent"	244 *81.3%*	16 *5.3%*	12 *4.0%*	9 *3.0%*	19 *6.3%*	300 *100%*
Rehearsed Fluent	263 *87.7%*	19 *6.3%*	9 *3.0%*	6 *2.0%*	3 *1.0%*	300 *100%*
Marginal Totals	821 *68.4%*	60 *5.0%*	47 *3.9%*	57 *4.8%*	215 *17.9%*	1200 *100%*

Table 4.11. Experiment 2: disfluency judgement distribution by stimulus type: crucial word.

the controls.

The distribution of 1-5 disfluency judgements at the crucial point for the four stimulus sets differed significantly ($\chi^2 = 677.296$, $df = 12$, $p < 0.0001$). The differences lay in the areas of distribution expected under the experimental hypothesis, with many fewer "fluent" judgements for the disfluent stimuli than for the controls (14% versus an average of 86.6%) and many more "4" and "5" judgements (12.3% and 62.0% versus 2.2% and 3.2%) (table 4.11 and figure 4.6).

A nonparametric analysis of variance (Friedman test) compared disfluency judgements at the crucial point in the four stimulus types for all subjects and all materials ($N = 300$). The differences in rank totals were found to be highly significant ($Xr^2 = 334.81$, $df = 3$, $p < 0.0001$), with the rank total for the spontaneous disfluent stimuli being much higher than those for the controls. No significant difference in rank totals was found between the three sets of controls.

To examine the differences between cells more closely, parametric analyses of variance (ANOVAs) were used and the data were treated as interval data. As for

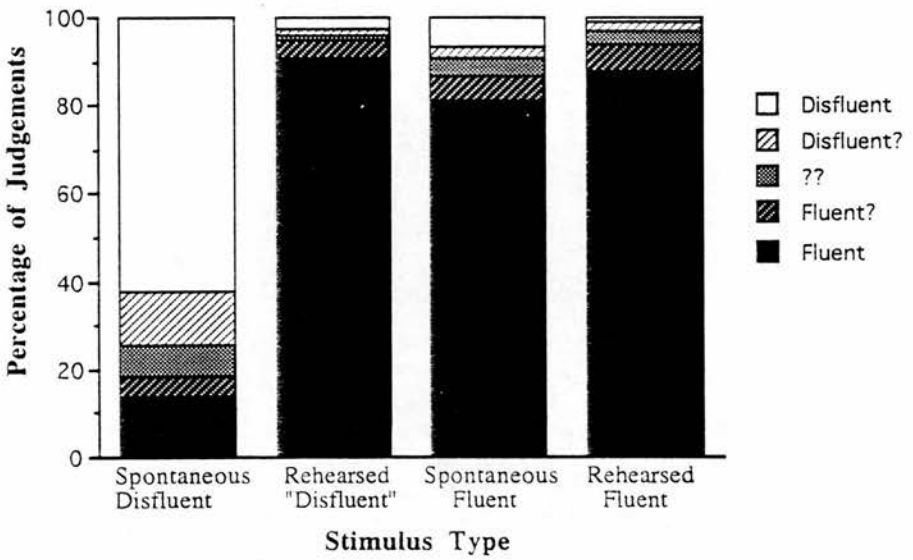

Figure 4.6. Experiment 2:fluency judgement distribution for first word of repair by stimulus type

Experiment One, cells for each crucial point were made up of totals of judgements for each stimulus type, by subjects and by materials.

Two-way ANOVAs with repeated measures for stimulus type were performed by subjects and by materials, with **fluency** (disfluent vs fluent) and **mode** (spontaneous vs rehearsed) as independent variables. Highly significant main effects both by subjects and by materials were found for **fluency** ($F_{1(1,19)} = 218.19$, $p < 0.0001$; $F_{2(1,19)} = 206.30$, $p < 0.0001$; $MinF'_{(1,46)} = 106.04$, $p < .0001$), for **mode** ($F_{1(1,29)} = 311.37$, $p < 0.0001$; $F_{2(1,29)} = 170.33$, $p < 0.0001$; $MinF'_{(1,47)} = 110.1$, $p < 0.0001$) and for the interaction of **fluency by mode** ($F_{1(1,19)} = 279.10$, $p < 0.0001$; $F_{2(1,29)} = 85.25$, $p < 0.0001$; $MinF'_{(1,43)} = 65.3$, $p < 0.0001$).

The cell means, illustrated in figure 4.7, strongly suggest that the effects and interaction were mainly caused by the high scores for spontaneous disfluent stimuli ($\bar{X} = 4.04$), rather than by any differences between scores for any of the control stimuli, whose means varied only slightly (from 1.21 to 1.39). *Post hoc* (Scheffé) tests with cell means from by subjects and by materials analyses confirmed this observation, showing that the fluency and mode main effects and the interaction were all caused by the scores for the spontaneous disfluent stimuli being significantly higher than for any of the control sets ($p < 0.01$) and that there were no significant effects of differences between judgements for any of the controls ($P > 0.05$).

In Experiment One, it was found that certain stimuli contained cues which might be of help to subjects in detecting oncoming disfluency. The same cues may have aided subjects in this experiment. In order to find out if they did, disfluency judgements were compared for stimuli with and without cues (as defined on page 87). As in Experiment One, there were 18 stimuli which contained cues and 12 which did not.

If cues had an effect on disfluency judgements, spontaneous disfluent stimuli with these cues would be expected to attract greater certainty than spontaneous disfluent stimuli lacking cues. However, the distribution of disfluency judgements in this experiment showed no effect of the presence of cues: the distribution of 1-5 judgements did not differ significantly between the with-cue and the no-cue conditions for the crucial word in spontaneous disfluent stimuli ($\chi^2 = 5.595$, $df = 4$, $p = 0.2315$) (table 4.12).

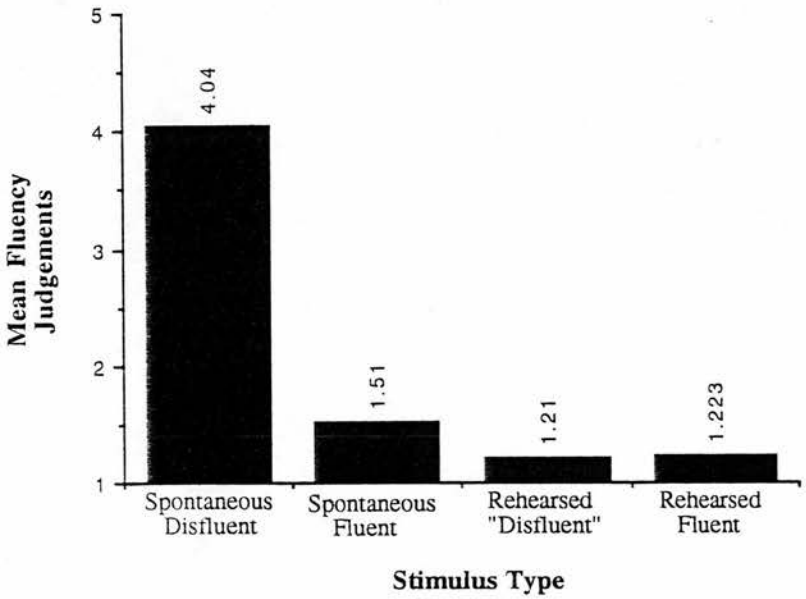

Figure 4.7. Experiment Two: Means of disfluency judgements at crucial word.

Cue ?	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Absent	17	3	5	18	77	120
%	*14.2*	*2.5*	*4.2*	*15.0*	*64.2*	*100*
Present	25	10	17	19	109	180
%	*13.9*	*5.6*	*9.4*	*10.6*	*60.6*	*100*
Marginal Totals	42	13	22	37	186	300
	14.0	*2.5*	*4.2*	*15.0*	*64.2*	*100*

Table 4.12. Experiment 2: disfluency judgement distribution for crucial word in spontaneous disfluent stimuli with and without pause or broken word (cue).

Analyses of variance were performed by subjects, as a three-way ANOVA with repeated measures for the binary conditions of fluency, mode and cue, and by materials, as a two-way ANOVA with repeated measures for fluency and mode, with cue as a grouping factor. A significant main effect of the presence of **cue** was found in the by-subjects analysis but not in the by-materials analysis ($F_{1(1,19)} = 24.77$, $p = 0.0001$; $F_{2(1,28)} = 2.37$, $p = 0.1349$). As expected, given the results of the previous ANOVAs, other main effects were highly significant in both by-subjects and by-materials analyses (**Fluency**: $F_{1(1,19)} = 243.90$, $p < 0.0001$; $F_{2(1,28)} = 191.28$, $p < 0.0001$; $MinF'_{(1,46)} = 107.267$, $p < 0.01$; **Mode**: $F_{1(1,19)} = 317.62$, $p < 0.0001$; $F_{2(1,28)} = 176.05$, $p < 0.0001$; $MinF'_{(1,46)} = 113.268, p < 0.01$), as was the **fluency by mode** interaction ($F_{1(1,19)} = 341.26$, $p < 0.0001$; $F_{2(1,28)} = 90.09$, $p < 0.0001$; $MinF'_{(1,40)} = 71.274$, $p < 0.01$). There were no significant interactions between cue and fluency or mode.

So the presence of cues at the interruption point did not affect the level of scores for the crucial word in spontaneous disfluent stimuli: no difference in mean scores was found between the with-cue and no-cue conditions.

A further analysis examined the effect of cues on disfluency judgements at the gate *prior* to the crucial word. In Experiment One, this gate contained the crucial word and the presence of a cue was found to result in higher scores, with the absence of cues leading to predicted-disfluency scores which were not different from those for the fluent controls. One possible explanation for the results of Experiment One was that subjects responded to hearing *disfluency* by giving more "disfluent" judgements and that the cues were seen as constituting disfluencies in themselves, rather than as cues to *oncoming* disfluency. If this was the case, then we would expect to find more judgements of "disfluent" in Experiment Two for the word prior to the crucial word for the stimuli which contained cues than for those with no cue.

The distribution of 1-5 judgements differed significantly between the with-cue and the no-cue conditions for the word prior to the crucial word in spontaneous disfluent stimuli ($\chi^2 = 13.867$, $df = 4$, $p = 0.0077$). The greatest differences lay in the number of judgements of "fluent", with more "1" judgements in the no-cue condition (80.8%) than the with-cue condition (62.2%) and in judgements of "don't know", where more occurred in the with-cue condition (10.6%) than in

Cue ?	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Absent	97	7	3	6	7	120
%	*80.8*	*5.8*	*2.5*	*5.0*	*5.8*	*100*
Present	112	13	19	17	19	180
%	*62.2*	*7.2*	*10.6*	*9.4*	*10.6*	*100*
Marginal Totals	209	20	22	23	26	300
	69.7	*6.7*	*7.3*	*7.7*	*8.7*	*100*

Table 4.13. Experiment 2: disfluency judgement distribution in spontaneous disfluent stimuli by presence of cue for word prior to crucial word.

the no-cue condition (2.5%) (Table 4.13).

As was done for the crucial word, ANOVAs were performed to examine the effects of cue, fluency and mode on disfluency judgements for the word prior to the crucial word (by subjects: a three-way ANOVA with repeated measures; by materials: a two-way ANOVA with repeated measures for fluency and mode, with cue as a grouping factor). There was no significant effect of **cue** in either analysis. A significant main effect of **fluency** was found in the by-subjects analysis ($F_{1(1,19)} = 6.68$, $p = 0.0182$) but not in the by-materials analysis. A significant main effect of **mode** was found in both analyses ($F_{1(1,19)} = 21.01$, $p = 0.0002$; $F_{2(1,28)} = 15.14$, $p = 0.0006$; $MinF'_{1,46} = 8.79$, $p < 0.01$). Significant interactions were found in the by-subjects analysis for **cue by fluency** ($F_{1(1,19)} = 17.67$, $p = 0.0005$), for **cue by mode** ($F_{1(1,19)} = 7.38$, $p = 0.0137$) and for **fluency by mode** ($F_{1(1,19)} = 30.77$, $p < 0.0001$), but none of these interactions reached significance in the by-materials analysis.

The distribution of disfluency judgements and the finding of interactions with cue in the ANOVA by subjects suggested that the presence of a cue had some effect on subjects' responses, with greater uncertainty as to the fluency of the stimuli which contained cues, but the effect was only weak and the mean disfluency judgement for these cases ($\bar{X} = 1.96$), while higher than those for the

no-cue condition and for the controls, was still within the region of "fluent" judgements. We conclude that listeners did not use pauses and broken words to indicate disfluency.

Finally, disfluency judgements for the crucial word in each individual spontaneous stimulus were compared with those for the spontaneous fluent controls. Judgements for the disfluent stimuli were found to be significantly ($p < 0.05$) higher than those in the fluent controls in 29 of the 30 cases (Wilcoxon signed rank tests). In the one case which did not produce a significant difference, six subjects judged the spontaneous fluent control to be disfluent as the speaker had stuttered slightly on the crucial word: the rehearsed version of the stimulus produced no "disfluent" judgements and differed significantly from the spontaneous disfluent version ($W = 0$, $N = 7$, $p = 0.0156$). The early identification of disfluency was thus possible in all cases tested in the experiment, which had been selected as a representative sample of the types of disfluency found in the corpus.

4.2.3 Results II: Word Recognition

In this experiment, as in Experiment One, subjects performed a word-recognition task at the same time as the fluency-judgement task. For Experiment One word recognition outcomes and disfluency judgements were compared to test the hypothesis that subjects gave more judgements of "disfluent" when they were unable to recognise the word they had just heard, but no such effect was found. Two hypotheses regarding the recognition of words on either side of disfluent interruptions were supported by the results: first, complete words before the interruption were found to be more likely to be missed than similar words in the control stimuli, while the percentage of immediate recognitions did not differ; second, fewer immediate recognitions occurred for the word following the interruption compared to words at a similar serial position in the controls. The results of Experiment Two were tested in the same ways and compared to those of the first experiment.

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	2115	2065	1810	2076	8066
%	*80.7%*	*85.7%*	*83.4%*	*87.6%*	*84.3%*
Late	328	274	286	243	1131
%	*12.5%*	*11.4%*	*13.2%*	*10.3%*	*11.8%*
Missed	177	71	74	51	373
%	*6.8%*	*2.9%*	*3.4%*	*2.2%*	*3.9%*
Total	2620	2410	2170	2370	9570

Table 4.14. Experiment 2: Word recognition outcomes for all words in all stimuli.

Overall word recognition performance

As in Experiment One, the total number of words presented in this experiment was 957 and each word token was presented to 10 subjects, resulting in a total of 9570 recognition outcomes. Each recognition outcome was classified as "immediate", where the word was recognised correctly on its first presentation, "late", where the word was recognised on a subsequent presentation, or "missed", where the word remained unrecognised. In assessing subjects' responses, the original word and its homophones, with or without correct inflection, were scored as correct recognitions.

The distribution of word recognition outcomes differed significantly between stimulus types, with fewer immediate and more missed recognitions in the spontaneous disfluent stimuli than in any of the controls ($\chi^2 = 97.957$, $df = 6$, $p < 0.0001$) (table 4.14). Within the controls, the distribution of outcomes for rehearsed "disfluent" stimuli differed significantly from the other controls, yielding more late recognitions ($\chi^2 = 17.39$, $df = 4$, $p = 0.0016$). These two results match those found in Experiment One.

Effect of failed word recognition on disfluency judgements

A possible artefact in the distribution of disfluency judgements was that subjects might have responded to not being able to recognise words by giving judgements of "disfluent". If words were more difficult to recognise around a disfluent interruption, this might affect our interpretation of the overall results of the experiment.

To test for this effect, the distribution of disfluency judgements for words recognised on first presentation was compared with that of other words. If an effect was present, we would expect to find a majority of "missed" words receiving "disfluent" judgements. If there was no effect at all, we would expect no difference between the distributions.

For these analyses, word recognition outcomes were classed as either "correct" or "missed" and disfluency judgements were assigned three categories: "fluent" (combining "1" and "2"), "don't know" ("3") and "disfluent" (combining "4" and "5").

In the first analysis, all word recognition outcomes and all

disfluency judgements were compared ($N = 9570$). The distribution of disfluency judgements differed greatly between "correct" and "missed" recognition outcomes ($\chi^2 = 850.9$, $df = 2$, $p < 0.0001$) (table 4.15). Subjects gave more "fluent" judgements for words they had recognised than for words not recognised: a large majority (86.4%) of "correct" outcomes coincided with "fluent" judgements; a smaller majority of "missed" outcomes (58.0%) also coincided with "fluent" judgements and the remaining "missed" outcomes were divided evenly (21%) between "don't know" and "disfluent". Subjects displayed greater certainty in their disfluency judgements when they had recognised the word: "don't know" judgements, were more frequent for "missed" words (21%) than for "correct" words (3.7%). While there were fewer "fluent" judgements on words which were not recognised on first presentation, the majority of "disfluent" judgements (71.8%) were still given to words which *had* been recognised: the main cause of responses of "disfluent" could not be said to be the non-recognition of the current word, since, if this were this case, we would expect a much larger proportion of "missed" recognitions to have resulted in "disfluent" judgements.

Recognition Outcome	Fluency Judgement			Totals
	Fluent	Don't Know	Disfluent	
Immediate	6965	297	804	8066
%	*88.9*	*48.5*	*71.8*	*84.3*
Missed	872	316	316	1504
%	*11.1*	*51.5*	*28.2*	*15.7*
Total	7837	613	1120	9570

Table 4.15. Experiment 2: Distribution of word recognition outcomes by disfluency judgements for all words in all stimuli.

Since these results may have been biased by the presence of disfluency in the spontaneous disfluent stimuli (the presence of disfluencies may have resulted in more "missed" outcomes), a second analysis was performed, examining only results from the three sets of control stimuli ($N = 6950$). The result was very similar to the first analysis, with a considerable difference between the distributions of disfluency judgements by word recognition outcomes ($\chi^2 = 646.38$, $df = 2$, $p < 0.00001$) (table 4.16). Removing spontaneous disfluent stimuli from the analysis made no significant difference to the distribution of disfluency judgements for "missed" words ($\chi^2 = 4.910$, $df = 2$, $p = 0.0859$) but did (predictably, given that the disfluent stimuli were excluded) result in a different distribution of outcomes for words recognised immediately, with fewer "disfluent" and more "fluent" judgements than in the first analysis ($\chi^2 = 46.919$, $df = 2$, $p < 0.0001$).

To summarise, it was found that if subjects failed to recognise the current word they were more likely to be uncertain about the fluency of the stimulus or to judge it disfluent than they were if they had recognised the word. But the majority of non-recognised words still received "fluent" judgements and the majority of "disfluent" judgements occurred where the word had been recognised. Non-recognition of the current word had some effect on disfluency judgements but could not be seen as the overriding factor in subjects' judgements of "disfluent".

These results match closely those found for Experiment One.

Recognition Outcome	Fluency Judgement			Totals
	Fluent	Don't Know	Disfluent	
Immediate	5351	197	403	5951
%	*89.7*	*48.8*	*69.6*	*85.6*
Missed	616	207	176	999
%	*10.3*	*51.2*	*30.4*	*14.4*
Total	5967	404	579	6950

Table 4.16. Experiment 2: Distribution of word recognition outcomes by disfluency judgements for all words in control stimuli.

Effect of disfluency on word recognition

In Experiment One it was found that the presence of disfluency had an effect on the recognition of the two words on either side of the interruption: the word before the disfluent interruption was recognised late less often and missed more often than similar words in the controls; the word following the interruption was recognised on first presentation less often than similar words at the same serial position in the utterance in the controls. The same analyses were performed on the data from this experiment.

The first analyses sought to confirm the finding from Experiment One that the word directly before a disfluent interruption was recognised late less often and missed more often than similar words in the controls. The distribution of word recognition outcomes for the word prior to the interruption in all spontaneous disfluent stimuli was compared to that for words at the equivalent place in the controls. A significant difference was found between the distributions ($\chi^2 = 57.633$, $df = 6$, $p < 0.0001$): fewer immediate recognitions and more late and missed recognitions were found in the spontaneous disfluent stimuli than in any of the controls (table 4.17). No significant difference was found between the three control sets. Eight of the 30 disfluent stimuli contained fragments as the word before the interruption. Since the correct recognition of fragments presents different problems from that of full words, it was decided to exclude fragments

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	234	275	270	283	1062
%	*78.0*	*91.7*	*90.0*	*94.3*	*88.5*
Late	38	17	24	16	95
%	*12.7*	*5.7*	*8.0*	*5.3*	*7.9*
Missed	28	8	6	1	43
%	*9.3*	*2.7*	*2.0*	*0.3*	*3.6*
Total	300	300	300	300	1200

Table 4.17. Experiment 2: Word recognition outcomes for word before disfluent interruption in all stimuli.

from the analysis of the recognition of the pre-interruption word and to focus on complete words. As was the case for Experiment One, the resulting distribution of word recognition outcomes still differed significantly between stimulus types, but at a lower level of significance and with a different pattern of outcomes in the spontaneous disfluent case ($\chi^2 = 12.61$, $df = 6$, $p = 0.0496$) (table 4.18): the main difference between the distribution of outcomes for spontaneous disfluent stimuli and that of the controls lay in a higher frequency of "missed" recognitions (5.2% of all outcomes compared to a mean of 2.03% for all controls); the frequency of immediate recognitions for the spontaneous disfluent stimuli did not differ significantly from the controls (89.1% compared to a mean of 90.57%), nor did that of late recognitions (5.7% compared to a mean of 7.4%). No difference was found between distributions of outcomes for the controls.

So the result of the analysis of word recognition outcomes for the word before the disfluent interruption was similar to that found in Experiment One, in that the word was recognised immediately as often as in the fluent controls but missed more often. Though the number of late recognitions was found to be lower in disfluent stimuli than in the controls in Experiment One, it was not significantly lower in this experiment.

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	205	208	204	213	830
%	*89.1*	*90.4*	*88.7*	*92.6*	*90.2*
Late	13	14	21	16	64
%	*5.7*	*6.1*	*9.1*	*7.0*	*7.0*
Missed	12	8	5	1	26
%	*5.2*	*3.5*	*2.2*	*0.4*	*2.8*
Total	230	230	230	230	920

Table 4.18. Experiment 2: Word recognition outcomes for word before disfluent interruption in all stimuli with complete words.

The second analyses compared the distribution of word recognition outcomes for the word following the interruption in spontaneous disfluent stimuli with that for equivalent words in the controls. In Experiment One it was found that this word was recognised on first presentation less often than the matched words in the controls. The results in this experiment were similar: the distribution outcomes differed significantly between stimulus types ($\chi^2 = 45.294$, $df = 6$, $p < 0.0001$) (table 4.19); as expected, a lower frequency of immediate recognitions (78.7%) was found in spontaneous disfluent stimuli than in the controls (average 90.9%). Within the three sets of controls, the spontaneous fluent set yielded more immediate and missed and fewer late recognitions than the rehearsed sets ($\chi^2 = 17.046$, $df = 4$, $p = 0.0019$).

4.2.4 Discussion

The main purpose of this experiment was to establish to a first approximation the point in an utterance at which it was possible for subjects to detect disfluency. In addition, the word recognition task which was performed at the same time as disfluency detection made it possible to test for effects of the recognition or

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	236	271	278	269	1054
%	*78.7*	*90.3*	*92.7*	*89.7*	*87.8*
Late	42	23	9	28	102
%	*14.0*	*7.7*	*3.0*	*9.3*	*8.5*
Missed	22	6	13	3	44
%	*7.3*	*2.0*	*4.3*	*1.0*	*3.7*
Total	300	300	300	300	1200

Table 4.19. Experiment 2: Word recognition outcomes for word after interruption in all stimuli.

non-recognition of words on disfluency judgements and effects of the presence of disfluency on word recognition in order to add further support to the results for the same task in Experiment One.

Detecting disfluency

The experiment was designed to test the hypothesis that disfluency could be recognised as early as the word directly following a disfluent interruption.

The results provided clear support for the experimental hypothesis. Disfluent stimuli overall and particularly at the crucial word yielded significantly more "disfluent" judgements than the control stimuli. The analyses of variance and post hoc tests showed that the mean disfluency judgements for the crucial word in spontaneous disfluent stimuli were significantly higher than those for the equivalent points in all the controls. The mean judgement of 4.04 for the crucial word in disfluent stimuli was a clear indication that subjects were able to detect disfluency with some confidence. The means for the controls, between 1.21 and 1.39, also show subjects' confidence in giving judgements of "fluent". This contrasts with

Experiment One, where subjects showed much less confidence in judging oncoming disfluency, with a mean of 2.55 for spontaneous disfluent stimuli (2.75 in the presence of cues), and also lower confidence in correct judgements of oncoming fluency (ranging from 1.82 to 2.04 for the crucial point in the controls).

In Experiment One, the presence of cues like pauses and uncompleted words were useful to listeners in detecting oncoming disfluency. In this experiment, where such cues were present, there was no significant effect on disfluency judgements for the crucial word. Only weak evidence was found for an effect of cue on judgements for the word before the interruption (the crucial word in Experiment One).

The experiment has relevance for models of speech understanding from both human and computational perspectives. Levelt (Levelt, 1983) identifies *recognition* as the first problem facing the listener when processing disfluent speech. The work described here represents an attempt to locate recognition points for disfluencies to a first approximation. The finding that listeners are able to detect the presence of disfluency within the first word of the continuation applies to a variety of types of disfluency, including some which are potentially still grammatical at the end of the crucial word (in this case, the word after the interruption), for example:

" they sent / a lot of their youngsters would go off ..."

"well in Edinburgh / no I think in Edinburgh ..."

Any model which relies primarily on syntactic cues for detection of disfluent speech would be unable to detect disfluency as early as the evidence presented here suggests is possible for the human processor: in the above examples, the earliest syntactic indications in the repair that the utterances are disfluent are found in the sixth and second words, respectively.

Having established that listeners are able to detect disfluency as early as the first word of the continuation, we are still left with the question of what linguistic or acoustic information makes this possible. In some cases there are obvious cues which were of use to subjects in Experiment One; in many cases syntactic information in the first word of the continuation was sufficient to inform subjects

that the utterance could not be fluent (assuming that the word had been correctly identified). Subsequent experiments examine the function of prosodic information in the process of disfluency detection and attempt to discover whether listeners are able to detect disfluency without having access to the syntactic information available when successful word recognition has taken place.

Word recognition and disfluency

The word recognition task allowed us a second opportunity to test the hypotheses examined in Experiment One. First, a possible artefact in the disfluency judgement task was that subjects may have simply responded to the inability to recognise the current word by giving higher scores for fluency, indicating uncertainty or even an assumption that disfluency was present. Second, it was hypothesised that the presence of disfluency would affect the recognition of the words on either side of a disfluent interruption in that the word before the interruption would be missed more often than similar words in fluent controls and the word immediately following the interruption would more likely to be recognised late than words in a similar serial position in fluent controls.

As in Experiment One, subjects showed a slight tendency towards uncertainty in their disfluency judgements when they had not recognised the word they had just heard. But the results do not support the hypothesis that disfluency recognition was due entirely to failure to recognise words in the vicinity of a disfluency: the majority of missed recognitions coincided with judgements of "fluent"; the majority of "disfluent" judgements coincided with immediate recognitions of words.

Complete words immediately before the interruption in spontaneous disfluent stimuli were recognised immediately as frequently as matched words in the controls but if they were not recognised on first presentation, they were usually missed altogether, rather than being recognised late: this is explained by the fact that in the fluent stimuli, right context provided information to help recognise words not identified on first presentation, whereas in the disfluent cases, the supporting right context was not present. Words immediately after the interruption

were recognised later and missed more often than words in the same serial position in the fluent controls: in these cases the words lacked the *left* context which aids recognition of words in mid-sentence. These results support the hypotheses and results from the same studies for Experiment One.

Chapter 5

Experiment 3: 35msec Gating Experiment

5.1 Introduction

Experiments One and Two have shown that listeners can usually detect disfluency by the offset of the first word of the fluent continuation. It was also found that the presence of disfluency coincided with more frequent failure of word recognition in the words immediately before the interruption and that words immediately after the interruption were recognised later than words at the same serial position in fluent stimuli, but that failure to recognise words did not lead to more judgements of "disfluent" in most cases.

These results beg several questions related to the detection of disfluency. The most immediate questions concern the point of recognition of disfluency and what information is required for a listener to be able to judge that a disfluency has occurred. One possibility, given the results of the first two experiments, where the word recognition task was performed very successfully even in disfluent stimuli, is that listeners are able to recognise that an utterance is disfluent simply by finding that a syntactic parse is impossible when the word following the interruption is recognised: this explanation is quite plausible, as only 3 of the 30 stimuli had continuations whose first word still allowed a possible parse. Another possibility is that acoustic information in the speech signal marking the presence of disfluency

is located within the word following the interruption but before the point where the listener can recognise the word.

To address this issue, an experiment was designed which allowed us to find more precise locations of detection points for both disfluency and for the words in the vicinity of a disfluency and thus to compare the two recognition points. In this way, it would be possible to determine whether it was necessary for a listener to have recognised the first word of the continuation before they could detect the disfluency or whether on the contrary there was sufficient non-syntactic information around the interruption point for the listener to detect that the utterance was disfluent, without identifying the word or its suitability in context.

To find recognition points for spoken words with greater precision than word-level gating, very short gates were used. It was decided to use increments of 35msec, this length being small enough to allow fairly precise location of any cues that listeners appeared to respond to. To limit the duration of the gating experiments we focussed on the portion of the utterance where disfluency was usually detectable: the two words on either side of the interruption point. Apart from the gate-size and the range of the gated section, the design of the experiment was similar to that of the first two studies: the gating method was used and subjects were asked to perform the two simultaneous tasks of word recognition and disfluency detection at each presentation.

The experimental design allowed several hypotheses to be tested.

The first analysis sought to confirm the finding of Experiment Two, that disfluency could be recognised within the first word of the continuation. As in Experiment Two, this hypothesis would be tested by comparing disfluency judgements at crucial points in disfluent stimuli with those at equivalent points in fluent control stimuli: if the hypothesis was to be supported, we would expect high rates of judgements of "disfluency" in the disfluent stimuli and no such judgements in the controls.

As regards the comparison between points of recognition of disfluency and the word following the interruption, there are three possible outcomes:

- under the **null hypothesis** there will be no significant difference between the recognition points of disfluency and those of the word following the disfluent interruption. The two recognition points will have no fixed order;

- under the **word-first hypothesis**, disfluency recognition will tend to follow word recognition;
- under the **disfluency-first hypothesis**, word recognition will tend to follow disfluency recognition.

In addition, in the previous experiments, word recognition performance was affected by the presence of disfluency: the word prior to a disfluent interruption, if not recognised on first presentation, was more likely to be missed than similar words in the controls; the word following a disfluency was less likely to be recognised on first presentation than words in similar serial position in the fluent controls. Presenting the words gradually in this experiment allowed us to examine word recognition results more closely.

5.2 Method

5.2.1 Materials

The materials used in this experiment were the same as those for the first two, except that in order to balance the materials over 4 groups of subjects, 2 of the original 30 test items and their controls were removed from the set. The materials thus consisted of:

- **Set A**: 28 spontaneous disfluent utterances;
- **Set B**: 28 rehearsed versions of "A" – rehearsed "disfluent" – with disfluency removed;
- **Set C**: 28 spontaneous fluent utterances, matched with "A" for structure and prosody;
- **Set D**: 28 rehearsed versions of "C".

In order to keep the running time for the experiment to a reasonable length, the materials were prepared for presentation to 4 subject groups, which were treated as matched quadruples. The sets of materials (spontaneous disfluent,

rehearsed disfluent, etc) were blocked by speaker, organised by latin square and then randomised to decide the order of presentation. The result was that each subject group heard 5 utterances from each of 4 speakers and 4 from each of 2 speakers. Each group heard a total of 7 items from each set of materials.

5.2.2 Subjects

Subjects were 43 native speakers of English, members of the University community (three groups of 11 and one of 10). The incentive of a small prize was offered for careful attention to the tasks.

5.2.3 Procedure

Before the experiment began, a taped introduction was given with full instructions and an example answer sheet was shown to the subjects. This was followed by a practice test consisting of three utterances produced by a speaker whose voice was not in the experiment proper.

The tape was then paused so that the practice test could be checked and so that subjects could ask questions.

Before the test items for each new speaker were presented, a short passage of conversation involving that speaker was heard, to help subjects get used to the voice. The test items were announced on the tape, giving the number for each item. 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. Gates were 35 msec long. Tones indicated to the subjects when the next presentation of a stimulus was about to begin: the tones were timed to allow subjects five seconds after each stimulus to write their responses and to precede the new stimulus by 2 seconds. A tone

and the announcement of the item number indicated the beginning of a new test item.

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

One line on the answer sheet was used for each new presentation of a stimulus. Subjects were asked to write down what they thought the latest word was at each gated presentation. They were asked to try to guess a whole word. Where they had already made a judgement and had not changed their mind on a subsequent presentation, they were asked to use ditto marks in the appropriate space on the answer sheet. Where they changed their mind about a previous judgement, they wrote the new judgement in the appropriate space without altering the original judgement. Where they could make no judgement, they put a horizontal line in the appropriate space.

The disfluency judgement task was the same as in the first experiment. 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.

A new answer sheet was used for each test item. The answer sheet had printed on it the item number, the context utterance and the beginning of the test utterance. Because this last information was available, subjects were better placed to devote attention to the detection of disfluency than they might be were they still trying to resolve the identity of earlier words.

There was space for answers to 50 presentations of each test item, each on a new line of the sheet, although the maximum number of actual presentations for any one item was 44. Each line provided spaces for up to three words to be written and ended with the printed numbers 1-5 for the disfluency judgement task. There was a reminder of what the numbers signified at the top of each column.

The structure of the presentation of test items for each speaker and required action by subjects commenced as in table 5.1.

There were a total of between 5 and 44 gated presentations per test item.

Subjects were tested in sound-proofed listening booths. The digital tapes

Tape:	"Speaker N"
Action:	None
Tape:	*Passage from conversation involving speaker N.*
Action:	Listen
Tape:	TONE.
	"Item one"
Action:	Listen
Tape:	*About 10 seconds of conversation prior to test utterance.*
	(Also printed on answer sheet)
Action:	Read and listen
Tape:	TONE.
	Beginning of test utterance up to point prior to gated section.
	(Also printed on answer sheet)
	(e.g. "this is the beginning of"...)
Action:	Tick text on answer sheet.
Tape:	TONE.
	Test utterance including 1st gate.
	(e.g. "this is the beginning of th-")
Action:	Attempt to guess word;
	disfluency judgement.
Tape:	TONE.
	Test utterance including 2nd gate.
	(e.g. "this is the beginning of the"
Action:	Attempt to guess word;
	disfluency judgement.
Tape:	TONE.
	Test utterance including 3rd gate.
	(e.g. "this is the beginning of the en-")
Action:	Attempt to guess word;
	disfluency judgement.

Table 5.1. Presentation method.

were played through headphones at a fixed amplitude.

The experiment was run in two 45 minute sessions, with a short break in between sessions.

5.3 Results I: disfluency judgements

These analyses seek first to confirm that disfluency can be detected before the end of the word following the interruption and then to compare points at which disfluency is detected with those at which the crucial word is recognised in order to test the "disfluency-first" hypothesis.

In the initial analysis, disfluency judgements for all four utterance-types are compared at gates before, at, and after the onset of the crucial word. For this analysis, we define a crucial reference point in the presentation of each stimulus (the onset of the word immediately following the interruption) and compare differences in disfluency judgements at points before and after this point for the different stimuli. To support the hypothesis that disfluency is detected by the end of the first word of the continuation, we expect to find many more "disfluent" judgements for gates within this crucial word in disfluent stimuli than for the equivalent gates in the fluent controls. We also expect to find differences within the disfluent stimuli between judgements before the onset of the crucial word and after the onset: judgements before the onset should be similar to those at equivalent points in the fluent stimuli.

In the second analysis, results are examined for the disfluent utterances only. This analysis tests the "disfluency or word first" hypotheses by comparing the point of disfluency recognition for each disfluent stimulus with the point of word recognition of the word following the interruption. For the null hypothesis, we expect to find no significant differences between gates of recognition for disfluency and the word following the interruption; for the "word first" hypothesis to be true, there should be significantly more cases where the gate at which the crucial word is recognised precedes that at which the disfluency is recognised; for the "disfluency first" hypothesis, we expect to find a significantly greater number of cases where the recognition of disfluency precedes the recognition of the crucial word.

An overview of the data revealed that for one disfluent stimulus, subjects appeared to have responded to a hesitation prior to the onset of the gating point, thus making an exception of the results for that stimulus. For this reason the responses for that stimulus and its related controls are disregarded and the number of stimuli is reduced to 4×27.

5.3.1 Disfluent vs Fluent stimuli: disfluency judgements

In Experiments One and Two, it was possible to compare disfluency judgements at specific points in the gated presentation, because these points were defined in units corresponding to whole words. It was not possible to do exactly the same analysis in this 35msec gating experiment. While spontaneous fluent controls were matched as closely as possible for structure and prosody with the disfluent stimuli, word lengths were inevitably different. Similarly, the rehearsed versions of the spontaneous stimuli were close, but not perfect, matches, in terms of word length: it is very difficult for speakers to imitate even their own speech rate with perfect precision. As a result of this, a straightforward gate-for-gate comparison of responses for the 4 sets of stimuli was not possible. But in order to compare disfluency judgements for different fluency conditions and at different points with respect to the onset of the continuation, it is useful to be able to identify equivalent points in all four sets of data. In order to achieve this, a simple method was devised to compare the data, based on a window of seven gates surrounding the onset of the second gated word, the middle (4th) gate being the one which contained the onset of the word, as determined from the waveform. In this way it was possible to compare responses for all four sets of stimuli at fixed temporal distances *before* and *after* the onset of the continuation, as well as *at* the gate which contained the onset.

With these 7-gate analysis windows selected, the data in this analysis thus consist of a total of 8127 disfluency judgements (43 subjects×27 stimuli ×7 gates).

The purpose of this analysis is to find out whether judgements for the disfluent stimuli differ significantly from those for the fluent controls. If the hypothesis is correct, there should be significantly more "disfluent" judgements for disfluent stimuli.

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Spontaneous Disfluent	663 *32.7%*	251 *12.4%*	297 *14.6%*	312 *15.3%*	507 *25.0%*	2030 *100%*
Spontaneous Fluent	1396 *68.8%*	326 *16.1%*	189 *9.3%*	63 *3.1%*	56 *2.8%*	2030 *100%*
Rehearsed "Disfluent"	1555 *76.3%*	284 *13.9%*	137 *6.7%*	35 *1.7%*	26 *1.3%*	2037 *100%*
Rehearsed Fluent	1426 *70.2%*	326 *16.1%*	178 *8.8%*	71 *3.5%*	29 *1.4%*	2030 *100%*
Marginal Totals	5040 *62.0%*	1187 *14.6%*	801 *9.9%*	481 *5.9%*	618 *7.6%*	8127 *100%*

Table 5.2. Experiment 3: disfluency judgement distribution within 7-gate analysis window by stimulus type

The distribution of disfluency judgements by stimulus type aggregated for all gates within the analysis window shows a clear difference between disfluent stimuli and the controls (table 5.2): disfluent stimuli have a lower proportion of "fluent" judgements and a higher proportion of "disfluent" judgements ($\chi^2 = 1961.95$, $df = 12$, $p < 0.0001$). There is also a difference between spontaneous fluent stimuli and both sets of rehearsed stimuli, with slightly fewer "fluent" judgements and slightly more "disfluent" in the spontaneous stimuli ($\chi^2 = 49.89$, $df = 8$, $p < 0.0001$).

Figures 5.1 to 5.4 show how the distribution of disfluency judgements varies at progressively later gates in the analysis window for all four conditions. Figure 5.1 illustrates that the distribution of judgements for the disfluent stimuli changes greatly over the course of the window. Where in the first gate there is a majority of "fluent" judgements, in the last gate there is a similar majority of "disfluent judgements. The distributions of judgements for the three control sets, on the other hand, do not change greatly over the course of the window, with a vast majority of "fluent" judgements and very few "disfluent" judgements at each of the 7 gates (figures 5.2– 5.4).

For the statistical analysis of the outcomes, three of the gates in the analysis window were selected: the first gate (three gates before the onset of the post-interruption word), the fourth gate (the gate containing the onset of the crucial word) and the seventh gate (three gates after the onset of the crucial word). We will refer to these three gates as "place 1", "place 2" and "place 3", henceforth. For each of these points, the mean of the disfluency judgements was used in the statistical analyses which follow. So, in by-subjects analyses, cells consisted of the mean judgement given by a subject for each condition and in the by-materials analyses, cells consisted of the mean judgement for each condition ($(total\ of\ judgements) \div (number\ of\ subjects)$) for a stimulus.

An initial inspection of the cells suggested two general patterns: disfluent stimuli received higher mean judgements than all three sets of controls at all three places in the analysis window; the mean disfluency judgements rose throughout the window for disfluent stimuli, but not for the fluent controls. The mean judgement at place 3 for each stimulus was higher in the disfluent condition than in each of the controls for every subject and for every stimulus (Wilcoxon signed

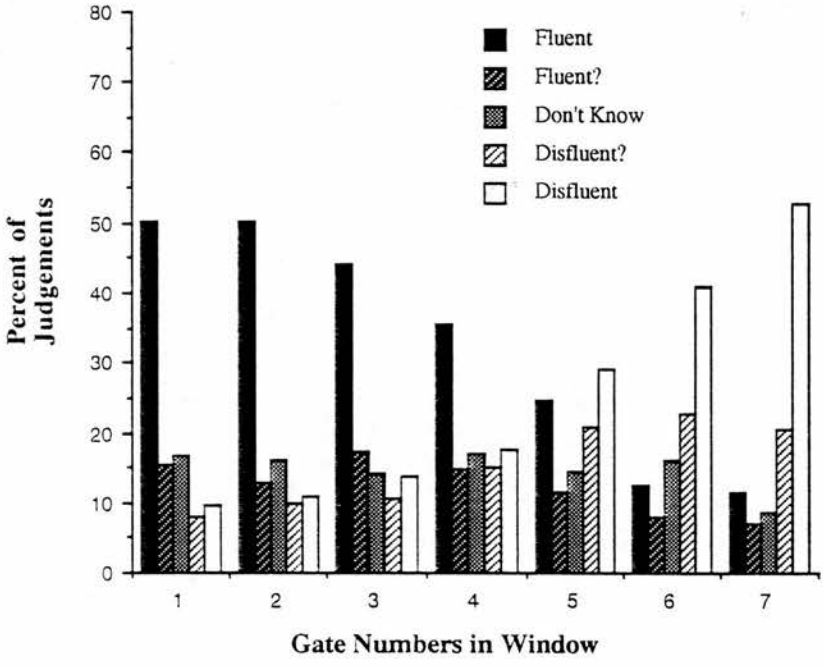

Figure 5.1. Experiment 3: Spontaneous Disfluent stimuli. Distribution of fluency judgements across 7-gate window, where gate 4 contains onset of first word of continuation.

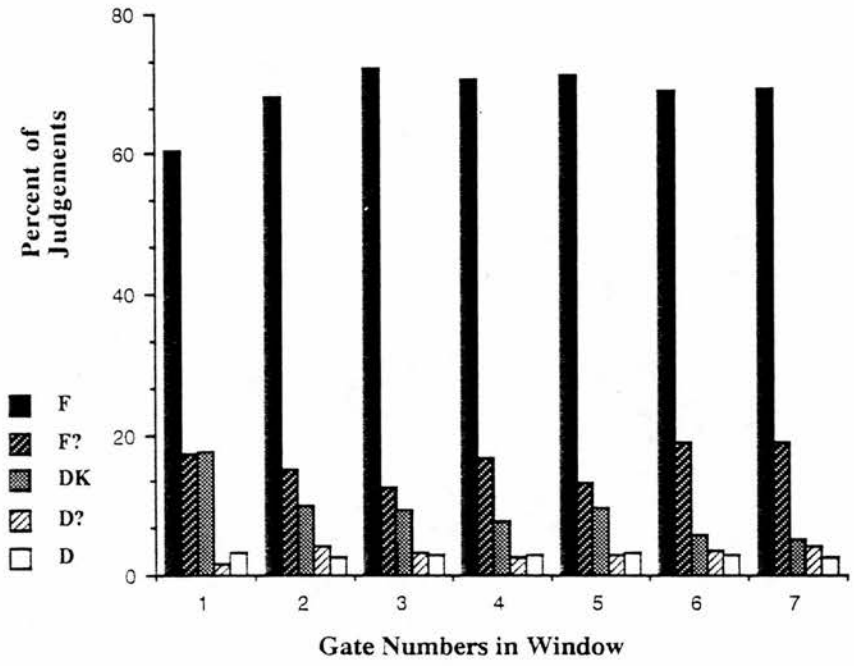

Figure 5.2. Experiment 3: Spontaneous Fluent stimuli. Distribution of fluency judgements across 7-gate window, where gate 4 contains onset of first word of continuation.

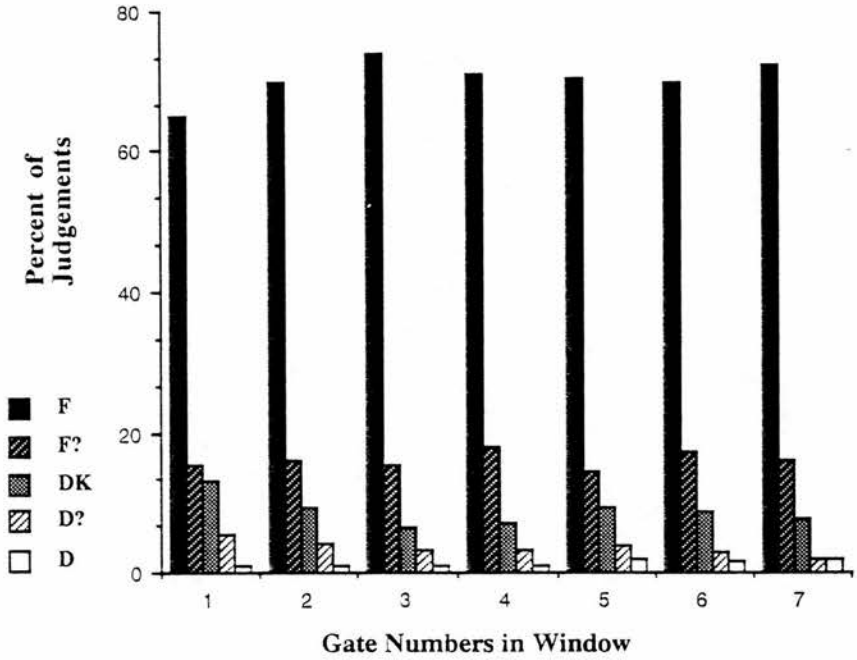

Figure 5.3. Experiment 3: Rehearsed "Disfluent" stimuli. Distribution of fluency judgements across 7-gate window, where gate 4 contains onset of first word of continuation.

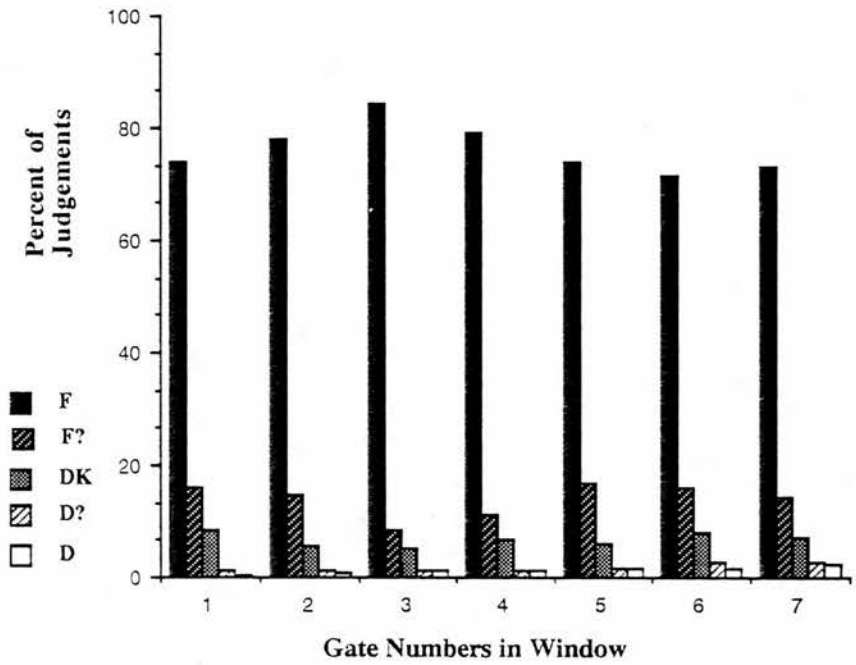

Figure 5.4. Experiment 3: Rehearsed Fluent stimuli. Distribution of fluency judgements across 7-gate window, where gate 4 contains onset of first word of continuation.

Variable	$\bar{X}_s$	sd_s	$\bar{X}_m$	sd_m
SD1	2.13	0.758	2.10	0.936
SD2	2.66	0.798	2.63	0.944
SD3	3.98	0.755	3.95	0.809
SF1	1.69	0.577	1.69	0.551
SF2	1.49	0.529	1.49	0.557
SF3	1.50	0.547	1.51	0.474
RD1	1.38	0.393	1.38	0.323
RD2	1.33	0.380	1.33	0.247
RD3	1.47	0.388	1.47	0.463
RF1	1.63	0.514	1.62	0.530
RF2	1.45	0.430	1.45	0.357
RF3	1.45	0.411	1.46	0.496

Table 5.3. Experiment 3: Cell means and standard deviations for 3-way ANOVAs (Mode by fluency by place), by subjects ($_s$) and by materials ($_m$). S = "Spontaneous", R = "Rehearsed", D = "Disfluent", F = "Fluent" and 1, 2 and 3 are places in analysis window, before onset, at onset and after onset of crucial word.

ranks test statistic (W) = 0, $N_{subjects}$ = 43, $N_{materials}$ = 27, $p < 0.0001$, for all three controls). Mean judgements at place 2 were also higher in the disfluent condition than in the controls for most subjects (minimum number in any of three controls= 41) and (minimum 23) stimuli ($p < 0.0001$) and at place 1 for most (minimum 35) subjects ($p < 0.0001$) and a majority (minimum 17) of stimuli. The mean judgement at place 3 in disfluent stimuli was usually higher than for place 2, and that for place 2, higher than for place 1 ($p < 0.0001$ both by subjects and by materials). Among the fluent stimuli, no similar pattern was found, the only significant differences between places in the windows being between place 1 and later places in the spontaneous fluent and rehearsed fluent sets, where these variables had a majority of *higher* values than their later-placed sisters. (These observations are reflected in the cell means displayed in table 5.3.)

More detailed analyses of the differences between judgements for disfluent and

Source	F_1	*df*	α	F_2	*df*	α	$MinF'$	*df*	α
Fluency	155.38	1,42	0.001	30.83	1,26	0.001	25.73	1,36	0.001
Mode	174.84	1,42	0.001	83.88	1,26	0.001	56.68	1,50	0.001
Place	55.76	2,84	0.001	37.43	2,52	0.001	22.40	2,114	0.001
F×M	200.04	1,42	0.001	65.87	1,26	0.001	49.55	1,43	0.001
F×P	189.86	2,84	0.001	63.13	2,52	0.001	47.38	2,86	0.001
M×P	109.66	2,84	0.001	36.01	2,52	0.001	27.11	2,86	0.001
F×M×P	85.59	2,84	0.001	34.26	2,52	0.001	24.47	2,93	0.001

Table 5.4. Experiment 3: F-ratios by subjects (F_1), by materials (F_2) and Minimum Quasi F-ratios ($MinF'$) for three-way ANOVAs with repeated measures for fluency, mode and place.

fluent and spontaneous and rehearsed stimuli over the 3-place analysis-window were carried out with three-way analyses of variance with repeated measures using means of disfluency judgements for the two **fluency** conditions, the two **mode** conditions (spontaneous and rehearsed) and three **place** conditions.

In both by-subject and by-materials ANOVAs, highly significant ($p < 0.0001$) main effects were found for all factors (**fluency**, **mode** and **place**) and all interactions (**F**×**M**, **F**×**P**, **M**×**P** and **F**×**M**×**P**). $MinF'$ was also found to be highly significant ($p < 0.001$) for all main effects and all interactions (tables 5.3 and 5.4).

In order to find out which differences between means contributed to the significant effect, a *post hoc* (Scheffé) test compared all 12 pairs of means for the fluency by mode by place interaction in the by-subjects and by-materials analyses. The most important question to be addressed in the test was whether the significance of the effects and interactions was caused just by differences between means for the spontaneous disfluent stimuli and the controls or whether there were also contributory differences between different conditions amongst the controls. The greatest differences between means were between all spontaneous disfluent cells and all control cells ($p < 0.01$, except for the difference between the first-placed spontaneous disfluent cell and the first-placed spontaneous fluent cell, where t'_{crit} was significant at $p < 0.05$). The means for the first-placed spontaneous disfluent cell (SD1) also differed significantly from the other spontaneous disfluent cells ($p < 0.05$ for difference with SD2 and $p < 0.01$ for difference with SD3). The

test confirmed the observations made on page 137, above, that there was very little difference between means for the control stimuli: the size of the differences between means of judgements for the 3 sets of control stimuli were smaller than t'_{crit} for $p < 0.05$ for both ANOVAs;

A possible reason for a hypothesised **place** effect, the rise in mean judgements from the first to the middle to the last gate, might be that uncertainty about the fluency or even the perception that the utterance was becoming disfluent might have increased simply when subjects heard the onset of a new word, especially given the unnatural task of listening to gated presentation of speech. If this were the case, differences between means for different gates in each set of controls would have been expected to contribute to the significance of the effect. This was not found to be the case: in the Scheffé test, only differences between means for different gates in the disfluent stimuli contributed to the significance, and not differences between means in any of the 3 sets of control stimuli.

The mean value for SD1 was close to the values for spontaneous fluent and rehearsed controls (table 5.3), but the Scheffé test suggested that the difference was large enough to contribute to the significance of the interaction, and the earlier Wilcoxon signed-rank test also showed significant differences between this variable and same-placed fluent controls. SD1 also differed from the fluent controls and was similar to the other disfluent variables in the size of its standard deviations (table 5.3), which were greater for all three spontaneous disfluent variables than for the controls, perhaps because of greater uncertainty about "disfluent" judgements than about "fluent" judgements. Since the gates which this value represents are *before* the onset of the word after the interruption, this suggests that there may be something in the signal before the crucial word begins which alerts listeners to the presence of disfluency. The most likely explanation for the higher mean judgements and the greater standard deviation for SD1 is that several (12) of the disfluent stimuli contained silent pauses at the interruption point.

It is also possible that the presence of pause was the *only* feature which induced judgements of "disfluent" (although the outcomes for individual stimuli suggest otherwise).

The pause effect

To test whether there was difference between judgements for stimuli with pause and those without, four-way ANOVAs were performed with repeated measures for **pause**, **mode**, **fluency** (2 levels each) and **place** (3 levels) in the by-subjects analysis and with one grouping factor (±pause) and the same three other factors in the by-materials analysis.

The cell means for the with-pause and no-pause conditions suggest that the presence of pause had an important effect on the overall results: the mean values for SD1 with no pause is at about the same level as the means for fluent controls; the means for SD1, SD2 and SD3 are higher in the with-pause condition than in the no-pause condition, but in the no-pause condition SD2 and SD3 are still higher than the fluent controls; the values for the controls do not differ between the pause conditions (since pauses only occurred in the disfluent stimuli, this was expected) (table 5.5).

The overall difference between means in the with-pause condition and means in the no-pause condition was found to be significant ($F_{1(1,42)} = 39.95$, $p < 0.0001$; $F_{2(1,25)} = 7.09$, $p = 0.0134$; $MinF'_{(1,34)} = 6.02$, $p < 0.025$). The **pause** by **mode** by **fluency** interaction was also significant ($F_{1(1,42)} = 11.64$, $p = 0.0014$; $F_{2(1,25)} = 8.85$, $p = 0.0064$; $MinF'_{(1,58)} = 5.03$, $p < 0.05$). Interactions between **pause** and **mode** and **pause** and **fluency** were significant in the by-subjects analysis (**Pse**×**M**: $F_{1(1,42)} = 5.21$, $p < 0.05$; **Pse**×**F**: $F_{1(1,42)} = 4.85$, $p < 0.05$), but not in the by-materials analysis. No other interactions with **pause** reached significance.

In conclusion, there was an overall effect of the presence of pause in disfluent stimuli. Where there was a pause, this was reflected in higher disfluency judgements before the onset of the continuation and continued higher judgements for the subsequent gates in disfluent stimuli. Where there was no pause in a disfluent stimulus, the disfluency judgement for the gate which preceded the onset of the continuation did not differ significantly from judgements at the same place in the fluent controls, but at the gate which contained the onset, the judgement was higher than in the controls and the judgements at the last gate clearly showed that subjects had identified disfluency without the aid of a silent pause. So,

	Pause				No Pause			
Variable	$\bar{X}_s$	sd_s	$\bar{X}_m$	sd_m	$\bar{X}_s$	sd_s	$\bar{X}_m$	sd_m
SD1	2.52	0.963	2.64	1.033	1.78	0.748	1.67	0.588
SD2	3.01	1.090	3.10	0.998	2.29	0.850	2.26	0.729
SD3	4.23	0.929	4.25	0.610	3.74	0.844	3.71	0.885
SF1	1.77	0.797	1.67	0.605	1.67	0.594	1.72	0.525
SF2	1.63	0.806	1.49	0.774	1.48	0.538	1.50	0.324
SF3	1.58	0.744	1.50	0.635	1.50	0.605	1.52	0.316
RD1	1.44	0.537	1.41	0.253	1.34	0.412	1.36	0.378
RD2	1.40	0.523	1.38	0.264	1.30	0.406	1.30	0.235
RD3	1.56	0.633	1.52	0.250	1.44	0.551	1.43	0.588
RF1	1.80	0.739	1.74	0.342	1.49	0.539	1.53	0.639
RF2	1.55	0.586	1.50	0.301	1.40	0.488	1.41	0.402
RF3	1.69	0.729	1.61	0.617	1.33	0.442	1.33	0.349

Table 5.5. Experiment 3: Cell means and standard deviations for 4-way ANOVAs (Pause by mode by fluency by place). S = "Spontaneous", R = "Rehearsed", D = "Disfluent", F = "Fluent" and 1, 2 and 3 are places in analysis window, before onset, at onset and after onset of crucial word.

the presence of pause in some disfluent stimuli caused greater uncertainty in disfluency judgements or earlier recognition of disfluency by some subjects, but, even in the absence of pause, subjects were still able to detect disfluency.

5.3.2 Disfluency detection vs word recognition

Having established that disfluent utterances are distinguishable from fluent utterances within the first word of the continuation, we now turn to the main question addressed in this experiment: which is recognised first – the disfluency or the word?

For this analysis, recognition of disfluency is judged to have been successful where subjects gave a judgement of "4" or "5". Word recognition was judged to be successful where subjects identified the correct word or a closely related word (e.g. "want" is taken as a correct recognition of "wanted", "was" [w@z] is accepted for "were" [w@] (in fast speech and with a non-rhotic accent)).

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 were 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. Under $\mathbf{H_0}$, there would be no difference between the gate at which the word and the disfluency are recognised; under the **disfluency-first** hypothesis, detection of disfluency would precede word recognition; under the **word-first** hypothesis, the crucial word would be recognised before the disfluency was detected.

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 disfluency judgements for one test item are disregarded from this point in the analysis for the reasons explained in section 5.3, page 130, above. The total number of cells is thus reduced to 290, since there were 11 subjects' responses for the stimulus in question. The resulting recognition rates are 246 (84.8%) for disfluency and 180 (62.1%) for the crucial word. A breakdown of the disfluency

and word recognition outcomes for all disfluent stimuli is shown in table 5.6 and illustrated in figure 5.5.

The gate number at which disfluency was recognised was compared with the gate number at which the first word after the interruption was recognised, for all cells of the remaining 27 disfluent stimuli. Disfluency recognition preceded word recognition in 192 (66.2%) of 290 cases. Word recognition preceded disfluency recognition in only 21 (7.2%) cases (and 5 of these were outcomes for one stimulus (see table 5.6). Word and disfluency recognition occurred at the same gate in 14.1% (41) of cases and in 12.4% (36) neither were recognised by the offset of the second word. These results showed that, overall, subjects recognised that the utterance was disfluent before they had recognised the crucial word. A matched t test was performed using only those cells where both disfluency and the crucial word were recognised by the end of the presentation of the stimulus (N=173) and the result was highly significant ($t = -9.53$, $df = 172$, $p < 0.0001$): on average, disfluency recognition preceded word recognition.

Further analyses examined the relationship between word onset and disfluency and word recognition. In 62 (21.38% of 290) cases, subjects identified disfluency *before the onset* of the word following the interruption; in 36 (12.41%) cases the disfluency was recognised at the gate which contained the onset of the word. The word following the interruption was never guessed before its onset, and was recognised at the gate which contained its onset in only 7 (2.41%) cases.

As can be seen in table 5.6, the distribution of outcomes differs between stimuli. In order to categorise the stimuli according to the distribution of outcomes, we will take a threshold level of 6 outcomes to define membership of a category: thus, if 6 or more subjects (i.e. a majority of subjects for any one stimulus) detected disfluency before the crucial word for any stimulus, that stimulus will be seen as a member of the set of stimuli where disfluency was recognised first. By this definition, the first fact to note is that disfluency was recognised by the end of the word following the interruption in 24 (88.9%) of the 27 stimuli. The crucial word was recognised in 17 (63%) stimuli. Disfluency was recognised before the crucial word in 21 (77.8%) cases, disfluency and the word were recognised at the same gate in 1 (3.7%) case and neither were recognised in 2 (7.4%) cases. Of the remaining 3 cases, 2 contain 5 "disfluency first" outcomes and 4 "same gate"

Item No.	D=W	D<W (no W)		W<D (no D)		Neither rec'd	Total
G1	10	1	(0)	0	(0)	0	11
G2	0	2	(0)	5	(3)	3	10
G3	0	6	(6)	1	(1)	4	11
G4	0	8	(7)	0	(0)	2	10
G5	3	8	(0)	0	(0)	0	11
H1	1	10	(1)	0	(0)	0	11
H2	0	7	(3)	2	(1)	1	10
H3	0	7	(4)	2	(2)	2	11
H4	1	10	(0)	0	(0)	0	11
J1	0	9	(8)	2	(1)	0	11
J2	0	8	(6)	0	(0)	2	10
J3	0	9	(0)	2	(0)	0	11
J4	4	6	(0)	1	(0)	0	11
J5	0	10	(5)	0	(0)	1	11
M1	2	7	(5)	1	(0)	1	11
M2	3	7	(0)	1	(0)	0	11
M3	4	5	(0)	0	(0)	2	11
M4	1	10	(1)	0	(0)	0	11
M5	4	5	(1)	1	(0)	0	10
N1	0	10	(10)	0	(0)	1	11
N2	1	7	(4)	1	(0)	2	11
N3	2	8	(1)	1	(0)	0	11
N4	3	7	(2)	0	(0)	0	10
P1	0	4	(3)	0	(0)	6	10
P2	0	3	(2)	0	(0)	8	11
P3	2	7	(4)	1	(0)	1	11
P5	0	11	(0)	0	(0)	0	11
TOTALS	41	192	(73)	21	(8)	36	290

Table 5.6. Comparison of recognition points of disfluency and word following interruption for each disfluent stimulus.

Where: Item No. = Item number of test stimulus;
D=W = Disfluency recognised at same gate as word;
D<W = Disfluency recognised before word;
(no W) = Disfluency recognised, but word not;
W<D = Word recognised before disfluency;
(no D) = Word recognised, but disfluency not;

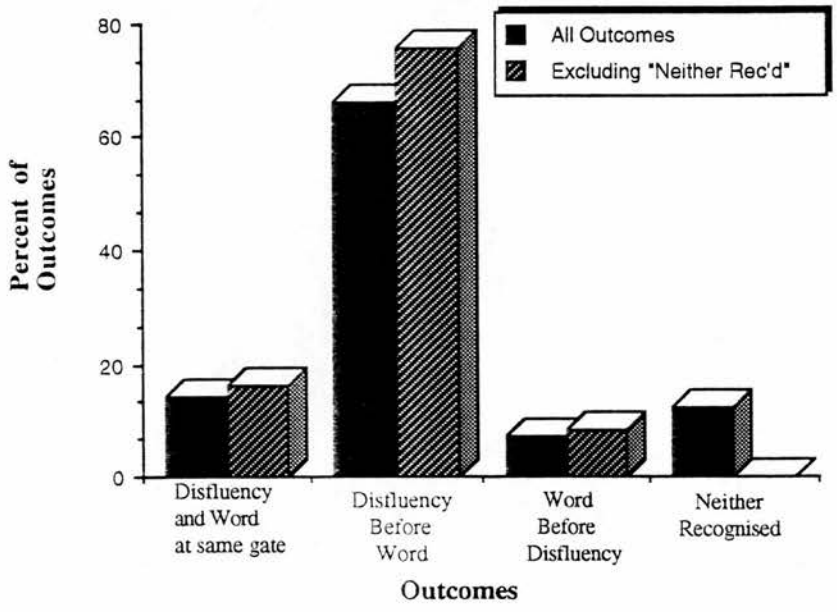

Figure 5.5. Experiment 3: Order of recognition of disfluency and word, for all outcomes and for all outcomes where one or other was recognised by offset of word after interruption.

outcomes and the remaining one contains 5 "word first" outcomes. Of the 21 cases where disfluency was recognised first, 4 contained a majority of responses indicating that the disfluency had been recognised before the onset of the word.

So the distribution of results for individual stimuli is similar to that found for the whole data set, in that disfluency recognition precedes word recognition in most cases and precedes the onset of the crucial word in some cases.

In conclusion, analysis of the disfluency judgements for disfluent stimuli confirmed the finding of Experiment One, that subjects could detect disfluency by the offset of the word following the interruption in most cases. Comparison of points at which disfluency was detected with points of word recognition showed that in most cases subjects detected disfluency in the signal before they had recognised the first word of the continuation. In no individual stimulus did a majority of subjects recognise the crucial word before they had detected the disfluency, although in one case this *was* achieved by 50% of the subjects. In one case both the word and the disfluency were recognised at the same gate.

5.4 Results II: word recognition

So far we have seen that subjects were able to detect disfluencies near the onset of the first word after the interruption and that disfluency was, in most cases, detectable before this word had been recognised. In contrast, in fluent controls, disfluency judgements did not change with respect to earlier "fluent" judgements at word boundaries which were matched with the disfluent word boundaries.

From these results, it seems likely that subjects were using cues other than lexical and syntactic in the detection of disfluency. However, a possible alternative explanation for the results might lie in subjects' using a strategy based on word recognition in making their disfluency judgements: it may be that words following disfluent interruptions were more difficult to recognise than words in fluent speech and that subjects gave judgements of "disfluent" simply because they could not recognise the word they were hearing. The combination of these factors and the effect of the presence of pause could produce precisely the results we have observed: subjects might have found more difficulty recognising words in disfluent stimuli and responded to their confusion by giving judgements of

"disfluent"; following recognition of the crucial word, subjects would then have been able to detect disfluency "legitimately" on the basis of lexical and syntactic information. If it is the case that failure to identify the current word elicited "disfluent" judgements, then the results of the analysis of the disfluency judgements suggest that subjects were able to recognise words very early in fluent stimuli – at the gate which contained the words' onset – since otherwise there would have been higher rates of "disfluent" judgements at these points than at the earlier or later points.

If, on the other hand, failure to recognise words did not prompt subjects to give "disfluent" judgements for fluent stimuli, it would also be important to show that subjects were able to correctly identify stimuli as *fluent*, in the absence of lexical and syntactic information.

So we now turn to the analysis of word recognition outcomes in this experiment in order to address the following questions:

1. did the presence of disfluency affect word recognition?

2. did the failure to recognise words prompt subjects to give more "disfluent" judgements in this experiment?

3. were fluent stimuli correctly labelled as "fluent" even before recognition of the current word?

At each gated presentation, subjects had been asked to try to guess the word or words being presented by writing a full word on the answersheet. In the analysis of this data, for comparisons of word recognition outcomes and disfluency judgements at individual gated presentations, the word recognition outcome at each gate was categorised as "correct" or "missed". "Correct" was assigned where the word was guessed correctly, or, in a very few cases, where a plausible and closely related word was identified (as in section 5.3.2). "Missed" was assigned in all other cases, including cases where an incorrect guess was made, or where the correct word was recognised later. In the analysis of word recognition outcomes for the word prior to the interruption, where the data are the outcomes at word-offset, a distinction is also made between "late" and "missed" recognitions, where "late" means that the word was recognised some time after the offset of the word.

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Missed	2498	1403	1795	1376	7072
%	*42.8*	*32.8*	*40.1*	*32.2*	*37.5*
Correct	3345	2877	2685	2903	11810
%	*57.2*	*67.2*	*59.9*	*67.8*	*62.5*
Total	5843	4280	4480	4279	18882

Table 5.7. Experiment 4: Overall word recognition outcomes for the four sets of stimuli (all gates, all outcomes).

This distinction is not possible for the word following the interruption, since there is no opportunity for subjects to recognise the word late in these cases.

5.4.1 Overall word-recognition performance

The entire experiment yielded a total of 18,882 word-recognition outcomes. The overall distribution of "missed" and "correct" outcomes differed significantly between the four stimulus types, with more "missed" outcomes in the "originally disfluent" (i.e. both spontaneous and rehearsed versions) than in the "fluent" sets ($\chi^2 = 174.226$, $df = 3$, $p < 0.0001$). The distribution of outcomes in the spontaneous fluent and rehearsed fluent sets (which contained the same set of words) did not differ significantly, but there *was* a significant difference between the distributions of outcomes in spontaneous disfluent and rehearsed "originally disfluent" stimuli, with slightly more "missed" judgements in the spontaneous cases ($\chi^2_y = 7.416$, $df = 1$, $p < 0.01$). These data are summarised in table 5.7.

5.4.2 The effect of disfluency on word recognition

If unintelligibility is the source of "disfluent" judgements, then the spontaneous disfluent utterances, which were often judged disfluent at very early gates, must have been quite hard to recognise. We should find that the word following the

interruption in disfluent stimuli yielded fewer correct responses than words at the equivalent points in the control stimuli. So before the effect of non-recognition of words on disfluency judgements is tested, we examine the hypothesis that words in the vicinity of disfluency are more difficult to recognise than similar words in fluent speech.

In Experiments One and Two, it was found that the words on either side of a disfluent interruption were affected by the presence of disfluency in different ways: the word before the interruption was recognised on first presentation as frequently as similar words in the controls, but if it wasn't recognised at this point, it was more likely to be missed entirely than similar words with coherent right contexts which allowed late recognition. The word after the interruption was recognised on first presentation less often than words in a similar serial position in the control stimuli. Similar results were expected for this experiment.

Results

We begin with the analysis of the distribution of word recognition outcomes for the word prior to the interruption (W_1), to test the hypothesis that word recognition outcomes for this word will differ between disfluent and fluent stimuli with respect to the frequency of late and missed recognitions.

Table 5.8 shows the distribution of word recognition outcomes for W_1 for the four different stimulus types. The distribution differs significantly over the four groups, the clearest difference being in the "late" and "missed" categories, where in the spontaneous disfluent stimuli there were fewer late recognitions and more missed recognitions than in the controls ($\chi^2 = 49.536$, $df = 6$, $p < 0.0001$). There was also a slight difference between the distribution of outcomes between the three control sets, with marginally more immediate recognitions in the rehearsed fluent set ($\chi^2 = 9.502$, $df = 4$, $p = 0.0497$).

The greater frequency of missed recognitions in the disfluent stimuli and the fact that the frequency of immediate recognitions did not differ significantly from the controls supports the hypothesis that the absence of textually coherent right context blocks the late recognition of words in disfluent speech, while the normal left context allows the same possibility of immediate recognition as for fluent stimuli.

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	183	250	247	265	945
%	*87.1%*	*89.3%*	*88.2%*	*94.6%*	*90.0%*
Late	2	23	21	11	57
%	*1.0%*	*8.2%*	*7.5%*	*3.9%*	*5.4%*
Missed	25	7	12	4	48
%	*11.9%*	*2.5%*	*4.3%*	*1.4%*	*4.6%*
Total	210	280	280	280	1050

Table 5.8. Experiment 4: Word recognition outcomes at word offset for the word prior to the interruption in spontaneous disfluent stimuli and the equivalent word in fluent controls.

Now we examine the distribution of recognition outcomes for the word following the disfluent interruption (W_2), testing the hypothesis that without coherent left context, disfluent stimuli suffer more missed recognitions than fluent stimuli where the left context contributes towards the word's predictability.

The distribution of "correct" and "missed" outcomes for W_2 in disfluent stimuli differs from that for the equivalent word in the controls ($\chi^2 = 87.134$, $df = 3$, $p < 0.0001$): immediate recognitions of W_2 occurred in 63.2% of disfluent stimuli as against an average of 88.1% in fluent stimuli. Comparisons with the distribution of recognition outcomes for W_1 (including "late" in the "missed" category for W_1) showed that the greater frequency of "missed" recognitions in W_2 was significant ($\chi^2_y = 34.033$, $df = 1$, $p < 0.0001$). For the other sets of stimuli, there was no difference between the distribution of outcomes between W_1 and W_2, except in the case of the rehearsed fluent set, where there were significantly more immediate recognitions for W_1 ($\chi^2_y = 5.329$, $df = 1$, $p = 0.021$). However, an overall analysis of all stimulus sets for both words omitting the outcomes for W_2 in the spontaneous disfluent set and combining the "late" and "missed" categories for W_1 shows no significant difference in the distribution of

Recognition Outcome	Spontaneous Disfluent	Spontaneous Fluent	Rehearsed "Disfluent"	Rehearsed Fluent	Total
Immediate	177	246	236	249	945
%	*63.2%*	*87.9%*	*87.4%*	*88.9%*	*81.8%*
Missed	103	34	34	31	202
%	*36.8%*	*12.6%*	*12.1%*	*11.1%*	*18.2%*
Total	280	280	270	280	1110

Table 5.9. Experiment 4: Word recognition outcomes at word offset for the word following the interruption in spontaneous disfluent stimuli and the equivalent word in fluent controls.

word recognition outcomes between "immediate" and "missed" for the 7 sets ($\chi^2 = 11.212,\ df = 6,\ p = 0.082$).

Perhaps word recognition outcomes for W_2 differed between disfluent and fluent stimuli because the set of words in the disfluent stimuli were of their nature more difficult to recognise immediately. For example, function words and shorter words are known to be more prone to late recognition than content and longer words (Bard *et al.*, 1988). Alternatively, it may have been possible that the presence of an incomplete preceding word in a quarter of the disfluent stimuli had a dramatic negative effect on the distribution of recognition outcomes for W_2, making it more difficult for subjects to segment the speech stream correctly. These possibilities were examined in further analyses.

There was indeed a greater proportion of function words among the disfluent W_2 stimuli (78.6%) than among the controls (49%). If the difference in recognition outcomes for W_2 between disfluent and fluent stimuli is solely due to the preponderance of function words in the disfluent stimuli and not due to the lack of left context, then we expect to find no difference in recognition outcome distributions for function words between disfluent and fluent stimuli. If, on the other hand, subjects recognised function words more easily in fluent than in disfluent stimuli, we might conclude that left context is a more important factor.

To test for the effect of word class, the distributions of recognition outcomes were further divided among responses to function or to content words. In disfluent stimuli for W_2, function words were recognised in 54.45% cases (number of outcomes= 220) and content words in 91.7% ($N = 60$). In fluent control stimuli (aggregated), function words were recognised in 87.1% of cases ($N = 410$) and content words in 89% ($N = 420$). The difference between recognition distributions for function words in disfluent and fluent stimuli was highly significant ($\chi^2_y = 78.820$, $df = 1$, $p < 0.0001$), showing that subjects were less able to recognise function words in disfluent stimuli than in fluent controls. The recognition distributions for *content* words in the two sets of stimuli did not differ significantly. The recognition distributions for function words differed significantly from those for content words in disfluent stimuli ($\chi^2_y = 25.05$, $df = 1$, $p < 0.0001$), with more "missed" outcomes for functions words, but not in fluent stimuli ($\chi^2_y = 0.593$, $df = 1$, $p = 0.44$).

So we conclude that the preponderance of function words in the disfluent stimuli was *not* the reason for the difference in recognition outcome distributions between fluent and disfluent stimuli, since there was no difference in outcomes between word classes within the fluent controls. There was, however a difference between recognition outcomes for content and function words in disfluent stimuli, the small number of content words being recognised as frequently as in fluent stimuli.

Another concomitant of intelligibility is millisecond length. Word length, measured in terms of the number of 35msec gates taken to present W_2, was compared for the disfluent set against all control sets. No significant difference was found in word lengths between any pair of word sets, the mean number of gates for disfluent W_2s (7.29) being only slightly less than for all fluent W_2s (8.29). Longer words were more likely to be recognised than shorter words amongst the disfluent stimuli ($r = -0.419$, $N = 28$, $p = 0.026$) but no such relationship was found for the fluent stimuli. Since the word lengths in disfluent and fluent stimuli did not differ significantly, word-length can not be seen as an important factor in the significance of the difference in distributions of recognition outcomes.

A final analysis compared cases where W_2 was preceded by an incomplete word with cases where W_1 was complete, to investigate the possibility that the

presence of a fragment preceding the word caused more "missed" word recognitions. "Immediate" recognitions occurred in 58.57% of cases ($N = 210$) where W_2 followed a complete word and 62.85% of cases ($N = 70$) where it followed a fragment. This difference was not significant. The presence of a fragment preceding W_2 clearly had no effect on subjects' ability to recognise the word.

This part of the analysis sought to find out whether the presence of disfluency had any effect on subjects' ability to recognise the words on either side of the interruption.

It was shown that complete words immediately prior to or ending in a disfluent interruption were recognised by their offsets as frequently as similar words in fluent controls. But when the word was not recognised by its offset, if it was in a fluent utterance, it was more likely to be recognised during the presentation of the following word than if it was in a disfluent utterance.

The word immediately following (or actually constituting) the disfluent interruption was recognised by its offset less frequently in disfluent stimuli than words following the matched point in fluent stimuli. It was concluded that this was due to the fact that in disfluent stimuli the word lacked a coherent left context, which was present in the fluent stimuli. Other possible explanations for this effect, such as a more frequent occurrence of function words or shorter words as first word of the continuation in the set of disfluent stimuli, or the presence of preceding fragments, were not found to have affected the result.

These results support the findings for Experiments One and Two.

5.4.3 The effect of failed word recognition on disfluency judgements

Having established that the word immediately following the interruption in disfluent stimuli was harder to recognise than the word after the matched point in fluent controls, we can now address the alternative explanation for the fact that subjects appeared to detect disfluencies before they had recognised the word following the interruption: the hypothesis is that subjects gave judgements of "disfluent" when they were unable to recognise the word they were hearing. Because the word following the interruption in disfluent stimuli was harder to recognise

than equivalent words in fluent stimuli, more "disfluent" judgements were given after the interruption in disfluent stimuli than at the equivalent points in fluent stimuli.

In order to test this hypothesis, it was necessary to compare the distribution of disfluency judgements across the 1-5 scale in cases where the current word had not been recognised with that when the word had been recognised. Since the results might have been biased by the inclusion of data from the disfluent stimuli if this alternative hypothesis were not correct, since, as we have seen, they included many cases where "disfluent" judgements were correctly given where the word had not yet been recognised and, in addition, they included many cases where both the word was recognised and disfluency was detected, only data from the fluent controls were used. If the hypothesis is correct, we expected to find that many more "disfluent" judgements occurred where subjects had not recognised the current word than where they had recognised it.

If the hypothesis is *not* correct, and subjects correctly identify disfluency on the basis of information available to them before word recognition, then we can also test the hypothesis that, as well being able to detect disfluency in the absence of lexical and syntactic information, subjects can correctly determine that an utterance is still *fluent* before they have the lexical and syntactic information to confirm this.

Results

Imminent disfluency judgements on the 1-5 scale and word recognition outcomes (±recognised) for every subject and every presentation were pooled for the three sets of fluent control stimuli (N= 13039).

The distribution of disfluency judgements differed significantly between points where the current word had been recognised correctly and points where it had not ($\chi^2 = 413.52$, $df = 4$, $p < 0.0001$), but, as can be seen from table 5.10, this difference was *not* due to there being substantially more "disfluent" judgements where the word was not recognised, but because there were more "don't know" judgements at these points: subjects gave "disfluent" judgements (4 or 5 on the scale) in 6.5% of cases where the word was not recognised and in 5.5% of cases where the word was recognised; "don't know" judgements were given in 15.4%

Recognition Outcome	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Word not Recognised	2812	762	703	165	132	4574
	61.5%	*16.7%*	*15.4%*	*3.6%*	*2.9%*	*100%*
Word Recognised	6367	1130	502	183	283	8465
	75.2%	*13.3%*	*5.9%*	*2.2%*	*3.3%*	*100%*
Marginal Totals	9179	1892	1205	348	415	13039
	70.4%	*14.5%*	*9.2%*	*2.7%*	*3.2%*	*100%*

Table 5.10. Experiment 4: distribution of disfluency judgements by word recognition outcomes in fluent stimuli for all gates and all subjects.

of cases where the word was not recognised and in 5.9% of cases where it was recognised; "fluent" judgements (1 or 2 on the scale) were given in 78.2% of cases where the word was not recognised and 88.5% of cases where it was. Within the "fluent" and "disfluent" categories, subjects showed greater certainty (more 1 or 5 than 2 or 4 judgements) where they had recognised the current word than where they had not.

On this evidence we reject the hypothesis that subjects gave judgements of "disfluent" when they were unable to recognise the current word: where subjects were unable to recognise the current word in fluent stimuli, they were more likely to give judgements indicating uncertainty (particularly "don't know") than where they had recognised the word, but no more likely to give "disfluent" judgements. A majority of judgements where the word was not recognised still fell into the "fluent" categories. We can thus reject the above hypothesis. We become more attached to the hypothesis that subjects could recognise that stimuli were fluent without having access to lexical and syntactic information.

5.5 Experiment 3a: Control Experiment

The results show that in the majority of cases disfluencies were detected before the word after the interruption had been recognised. One possible artefact in this

experiment was that the dual task of disfluency detection and word recognition may have distracted subjects from the word recognition task and thus delayed the point of recognition. If this was the case then the overall finding of Experiment Three might be put in doubt and it might not be possible to reject the hypothesis that subjects were just identifying disfluency on the basis of lexical and syntactic information.

To test for this possibility a second experiment was run with 35msec gates but with only the single task of word recognition. Materials were the stimuli used for one subject-block in Experiment Three, presented under the same conditions. The aim of the experiment was to determine whether word recognition latencies, measured in terms of the number of gated presentations needed by subjects to recognise words, differed between the original experiment where subjects had to perform two tasks and this one where there was only one task.

The experimental hypothesis was that under the single task condition subjects would correctly identify words sooner than under the dual task condition.

5.5.1 Method

Materials and Design

The experimental materials were identical to those used for the first subject-block in Experiment Three: a total of 28 stimuli, 7 from each set of stimuli (spontaneous disfluent, spontaneous fluent, rehearsed "disfluent" and rehearsed fluent), as described on page 125.

Subjects

Subjects were 10 native speakers of English, members of the University community. None reported any hearing disorders.

Procedure

Listening conditions were the same as for Experiment Three. Subjects were seated in sound-proofed booths and equipped with high-quality headphones and answersheets.

An introduction was read out, explaining the nature of the experiment and the task involved. The word recognition task was the same as that described in section 5.2.3. Subjects were warned that the speech they were to hear came from spontaneous conversations and would therefore sometimes contain disfluency.

The experiment then proceeded as for Experiment Three.

Answersheets were the same as for Experiment Three except that they contained only spaces for responses for the word recognition task.

The experiment was run in two 45 minute sessions, with a short break between sessions.

5.5.2 Results

For each of the 28 stimuli, there were two words to be identified. The recognition score for each word was determined on the basis of the number of gated presentations required before subjects could identify it. If a word was not recognised by its offset, it was given a score of the total number of gates for that word + 1. The mean recognition score for each word was calculated and the results compared to the recognition scores for the same words in Experiment Three.

Over all words in this experiment, recognition came on average 0.28 gates earlier than it had in the dual task experiment ($t = 3.03$, $df = 55$, $p = 0.0038$). For a dual-task "recognition delay" to affect the finding that disfluency was detected before W_2 had been identified, the effect should be found for W_2 in spontaneous disfluent stimuli. A comparison of the number of gates to recognition for W_2 in the seven disfluent stimuli retested here revealed no significant difference between the two experimental conditions ($t = 0.64$, $df = 6$, $p = 0.5446$).

5.5.3 Discussion

A control experiment was run to check for an effect of the dual task in Experiment Three on the outcome of the word recognition task. Subjects in this experiment were only asked to perform the word recognition task and did not have to make disfluency judgements. All the stimuli from one subject block used in Experiment Three were presented. The mean number of gates to recognition (across subjects) for each word presented was compared for the dual and single task conditions.

The results suggest that there was a significant effect of the dual task, with words being recognised slightly sooner overall when subjects had only the word recognition task to perform. But the difference between the mean number of gates to recognition for the two conditions was very small. Importantly from the point of view of the results of Experiment Three, no difference was found between the two conditions for the word after the interruption in disfluent stimuli. This means that for this section of the stimuli of Experiment Three, the dual task had no obvious effect on the overall finding that subjects could detect disfluency before recognising the word after the interruption.

5.6 Discussion

In Experiment Three subjects were presented presented with a sample of 28 spontaneous disfluent stimuli and three sets of fluent controls in 35msec gating format and required to make disfluency judgements and attempt to recognise the words they heard. The main question addressed was whether subjects required to have recognised the word following the interruption in disfluent stimuli before they could detect the disfluency. In addition, it was possible to use the data to repeat some of the analyses performed for Experiments One and Two, regarding both the detection of disfluency and word recognition. In this section we discuss the latter results before looking at the main findings.

Experiment Two showed that subjects could detect disfluency in an utterance by the offset of the word following the interruption. This finding was confirmed in the present experiment. But with the gradual presentation of words in 35msec gates in this experiment, it was possible to find earlier detection points for disfluencies. A significant rise in disfluency judgements was found as early as the gate which contained the onset of the word after the interruption in disfluent stimuli which did not contain a pause and before the onset in stimuli which contained a pause at the interruption point.

The experiment allowed examination of subjects' ability to recognise the two words on either side of the interruption. The results of the analysis of the recognition of these two words that were found for Experiments One and Two were confirmed here. The word before the interruption was recognised by its offset

as frequently as similar words in the fluent controls but was missed more often. Over 95% of the first gated words in the controls were recognised by the end of the following word, but only 88% of such words were recognised by the same point in disfluent stimuli: we assume that this is because the late recognition of words is aided by a syntactically and semantically coherent right context, which was present in fluent stimuli but not in disfluent stimuli. The word immediately following (or actually constituting) the disfluent interruption was recognised by its offset less frequently in disfluent stimuli than words following the matched point in fluent stimuli. It was concluded that this was due to the fact that in disfluent stimuli the word lacked a coherent left context, which was present in the fluent stimuli. Other possible explanations for this effect, such as a more frequent occurrence of function words or shorter words as first word of the continuation in the set of disfluent stimuli, or the presence of preceding fragments, were not found to have affected the result.

The non-recognition of a word did not lead to subjects judging the stimulus to be disfluent. There was a slight tendency towards uncertainty in disfluency judgements where the word had not been recognised but in the majority of cases subjects were able to make successful disfluency judgements whether or not the word had been recognised. It is interesting to note that it was still possible for subjects to judge correctly that a stimulus was fluent when the word had not been recognised. This finding supports and adds to the similar finding in Experiment Two.

The most important finding in this experiment concerned the relationship between the points at which disfluency was detected and the word following the interruption was recognised. In the majority of cases, subjects were able to detect disfluency before they had recognised the first word of the continuation. In some of these cases, disfluency was detected before the onset of the crucial word. This was found to be the effect of the presence of silent pauses longer than 130msec at the interruption point, which occurred in 12 of the 27 disfluent stimuli. For the 15 stimuli which contained no such pause, the mean disfluency judgement before the crucial word did not differ from the mean judgements in the controls, but the judgements at and after the onset of the word following the interruption for stimuli with no pause were still found to be significantly higher

than in the controls. So, the presence of pause was found to facilitate early detection of disfluency, but where there was no pause, subjects were still able to detect disfluency within the first three gates of the crucial word. A possible artefact was that word recognition may have been delayed by the complex nature of the dual tasks of word recognition and disfluency detection. This possiblity was tested in a control experiment where subjects were asked only to recognise the words and not to make disfluency judgements. In the single task condition, the mean recognition point was 0.28 gates earlier than in the main experiment. But no difference was found in the recognition points for the word after the interruption in disfluent stimuli. The control experiment did not challenge the main result and the "disfluency-first" hypothesis is not rejected.

We conclude that listeners can detect discontinuities in speech without having access to the syntactic and semantic information in the continuation which is available when word recognition has succeeded. This leaves us with the question of what information subjects used in making their judgements. One likely source is in the prosodic characteristics of disfluency. Darwin (1975, discussed in Chapter 2, page 33) showed that listeners make use of continuity in the intonation of fluent speech in following a particular signal. No previous studies have looked specifically either at the nature of any prosodic discontinuities that occur in disfluent speech, nor at the perceptual characteristics of the prosody of spontaneously produced disfluency, but the results of this experiment suggest that this information may play an important rôle in the understanding of such speech: if listeners pay attention to the prosodic continuity of speech, then they may be particularly sensitive to the occurrence of *discontinuity*. One aspect of prosodic continuity, timing, has already been found to be of use to subjects in Experiments One to Three: in the next experiments, we investigate the rôle of intonation.

Other cues that subjects may have used in Experiment Three lie on the acoustic and phonetic levels. In chapter 7, we look for such cues by matching responses in the 35msec gating experiments presented here and in Chapter 6 to waveforms, spectrograms and pitch tracks of the stimuli.

Chapter 6

Experiments 4 and 5: Detecting disfluency in low-pass filtered speech

6.1 Introduction

Experiments One, Two and Three have shown that listeners are able to detect disfluency in speech at an early stage and often even before having recognised the word following the interruption. This suggests that information is used in this task which is available sooner than the lexical information which would allow syntactic assessment of the discontinuity. A likely source of such information is in the prosodic characteristics of the speech surrounding the interruption.

Darwin (Darwin, 1975) has shown that listeners make use of prosodic continuity in attending to a source of fluent speech, even to the extent that prosodic information can temporarily override syntactic and semantic information. If the prosodic pattern of the speech a person is listening to is disrupted in some way, it is likely that such disruption will be detected very quickly (see Chapter 2). It has been suggested (e.g. most pertinently by Hindle (1983) and Levelt (1983)) that prosodic information may be of use in understanding disfluent speech. The experiments described in the previous chapters support this view. The next experiments directly address the question of whether prosodic information might

play a rôle in the processing of disfluent speech.

In order to examine listeners' responses to only the prosodic information in the stimuli, it was necessary to find a method of removing all segmental information, while still maintaining some degree of naturalness in the speech signal. Low-pass filtering makes it possible to remove higher sound frequencies and consequently the segmental information in the formants, while still maintaining the fundamental frequency (F_0) required for intonation to be perceived correctly. The auditory effect is similar to hearing speech coming from a neighbouring room through a wall: when the listener can hear that a person is speaking, can tell how quickly and with what emotional character the speech proceeds, but can not make out the words (see Chapter 2).

This chapter describes two experiments which use low-pass filtered speech to find out whether listeners can detect disfluency in speech from prosodic information alone. The first experiment presents whole utterances, low-pass filtered from the point of interruption to the end. The second experiment focuses on the two words on either side of the interruption. As in Experiment Three, the two words are presented with the 35ms gating technique: but in this experiment, the two words are also low-pass filtered, allowing subjects to base their judgements on prosodic information alone.

6.2 Experiment 4: Detecting disfluency in low-pass filtered speech

The question to be addressed in the first experiment with low-pass filtered speech is: given a normal onset to an utterance in spontaneous speech, can listeners judge, from prosodic information alone, whether the continuation is fluent or disfluent? To this end, a set of disfluent stimuli was selected (the same set as used in the previous experiments) and low-pass filtered from the interruption point to the end. A set of fluent stimuli taken from the same corpus and matched with the disfluent set for structure and prosody as far as possible (also the same set as used in the previous experiments) was treated in the same way, being low-pass filtered from the point equivalent to the interruption point in the disfluent

stimuli. With these materials, it was possible to address the question by eliciting disfluency judgements for sets of disfluent and fluent materials with similar onsets.

The experimental hypothesis was that listeners would be able to distinguish disfluent from fluent continuations on the basis of the prosodic information contained in the continuation, given all linguistic information up to the interruption point.

6.2.1 Method

Materials

Materials were the same spontaneous utterances as those used in the previous experiments: 30 matched pairs of disfluent and fluent utterances taken from a corpus of 6 spontaneous conversations, digitally recorded in a studio. The disfluent utterances had been selected as a representative sample of the distribution of disfluencies in the corpus as a whole. The fluent utterances had been selected to match the disfluent ones for structure and prosody as far as possible.

The stimuli were sampled at 20KHz and prepared for the experiment using ILS software on a MASSCOMP. Each stimulus was low-pass filtered from the point of disfluent interruption in the disfluent stimuli and the equivalent point in the fluent stimuli. The filter was adjusted individually for each of the six speakers to a level at which no formants were audible, but F_0 remained intact and intensity variations were still maintained. The filter used was a 5 pole Butterworth low-pass filter, designed using ILS software. The cutoff levels were decided individually for each speaker, the decision being based on the F_0 levels which occurred in the materials. Table 6.1 shows the maximum F_0 found in the materials for each speaker and the filter cutoff levels applied.

The 60 stimuli were presented in blocks of ten by speaker and randomised within each block.

A digital tape was prepared for the experiment. Each test item was preceded by a tone and presented three times in succession; on the first presentation, the test utterance was preceded by up to ten seconds of the conversation prior to it; on the second and third presentations, only the test utterance was heard; about five seconds of silence separated each presentation. The experiment proper was

Speaker	Sex	Max F_0 (Hz)	Filter Cutoff (Hz)
1	F	239.5	250
2	F	227.6	250
3	F	275.2	300
4	M	241.5	250
5	M	191.9	200
6	M	205.0	250

Table 6.1. Experiment 4: Maximum F_0 and Low-pass Filter Cutoff per Speaker.

preceded on the tape by a short practice session consisting of three test items using materials not included in the corpus.

An instruction sheet and answer sheets were prepared. The answer sheets were printed with speaker, item, and presentation numbers and included a line for each presentation consisting of the numbers 1 to 5, which subjects were to use to register their disfluency judgements.

Subjects

Subjects were 12 students from the Linguistics department of Edinburgh University, members of an honours and MSc class in speech technology and speech perception. None had taken part in previous experiments which used the same materials. All were native speakers of English and could be expected to be familiar with the range of accents represented in the materials. All reported having normal hearing.

Procedure

All 12 subjects heard the same set of materials.

Subjects were seated in individual listening booths and provided with an instruction sheet, answer sheets and high-quality headphones. They were told that at the point at which the low-pass filter was applied, some of the stimuli they would hear would continue fluently and some would continue disfluently, in that the speaker would repeat or change something that they had said. They were asked to listen carefully to each stimulus and to make a judgement as to its

fluency, using the 1-5 scale on the answer sheet (1 signified "fluent", 5, "disfluent", as in the previous experiments). They were warned that in some cases they might hear some form of disfluency in the speech before the low-pass filter was applied and that they should ignore this. They were also advised that they should not assume that the presence of a pause meant that the utterance continued disfluently.

Having read the instructions, subjects were invited to ask questions for clarification. A practice test consisting of three items followed, after which the practice answer sheets were checked and subjects allowed to ask more questions, if necessary.

The experiment was then run in two sessions of about 20 minutes each.

6.2.2 Results

Twelve subjects gave 3 judgements on a total of 60 test items, consisting of 30 fluent–disfluent pairs of stimuli, resulting in a total of 2,160 data points.

Subjects heard and responded to each item three times. The difference between distributions of disfluency judgements on each presentation suggested that subjects became more confident in their judgements after they had heard the same stimulus a second and third time. As figures 6.1 and 6.2 illustrate, the first presentation of both disfluent and fluent stimuli yielded more "don't know" judgements than later presentations and the last presentation yielded the lowest percentage of "don't know" and more "1" and "5" judgements than the earlier presentations. The differences between disfluency judgement distributions for the three presentations were significant for both disfluent ($\chi^2 = 80.015$, $df = 8$, $p < 0.0001$) and fluent stimuli ($\chi^2 = 59.029$, $df = 8$, $p < 0.0001$) (table 6.2).

Figures 6.1 and 6.2 also illustrate the result relevant to the experimental hypothesis: subjects gave more "disfluent" judgements for disfluent stimuli and more "fluent" judgements for fluent stimuli. Distributions for the same presentation differ significantly between disfluent and fluent stimuli for all three presentations (1st presentation: $\chi^2 = 113.565$, $df = 4$, $p < 0.0001$; 2nd presentation: $\chi^2 = 135.349$, $df = 4$, $p < 0.0001$; 3rd presentation: $\chi^2 = 144.966$, $df = 4$, $p < 0.0001$).

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Disfluent 1	56	63	87	82	72	360
%	*15.6*	*17.5*	*24.2*	*22.8*	*20.0*	*100*
Disfluent 2	72	54	28	107	99	360
%	*20.0*	*15.0*	*7.8*	*29.7*	*27.5*	*100*
Disfluent 3	85	37	32	76	130	360
%	*23.6*	*10.3*	*8.9*	*21.1*	*36.1*	*100*
Fluent 1	133	118	59	39	11	360
%	*36.9*	*32.8*	*16.4*	*10.8*	*3.1*	*100*
Fluent 2	182	86	28	46	18	360
%	*50.6*	*23.9*	*7.8*	*12.8*	*5.0*	*100*
Fluent 1	209	64	25	37	25	360
%	*58.1*	*17.8*	*6.9*	*10.3*	*6.9*	*100*
Marginal Totals	737	422	259	387	355	2160
	34.1%	*19.5%*	*12.0%*	*17.9%*	*16.4%*	*100%*

Table 6.2. Experiment 4: disfluency judgement distribution by fluency and presentation of stimulus.

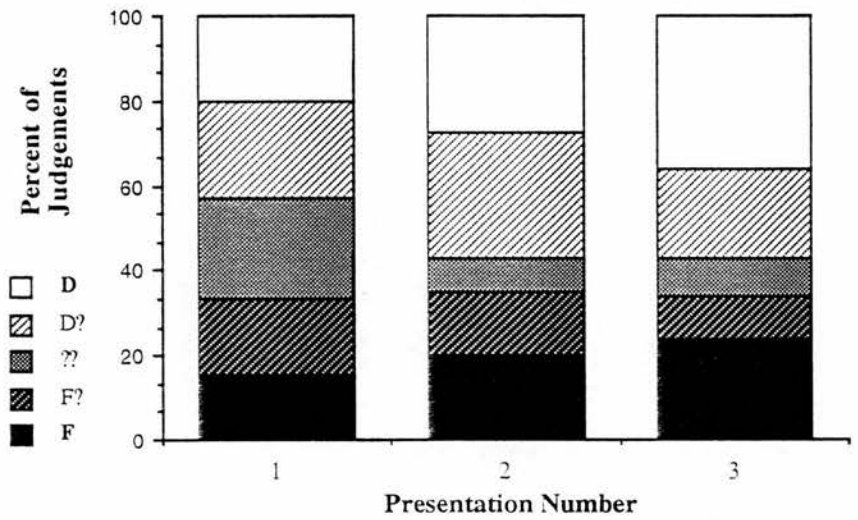

Figure 6.1. Experiment 4: Distribution of fluency judgements by presentation: disfluent stimuli.

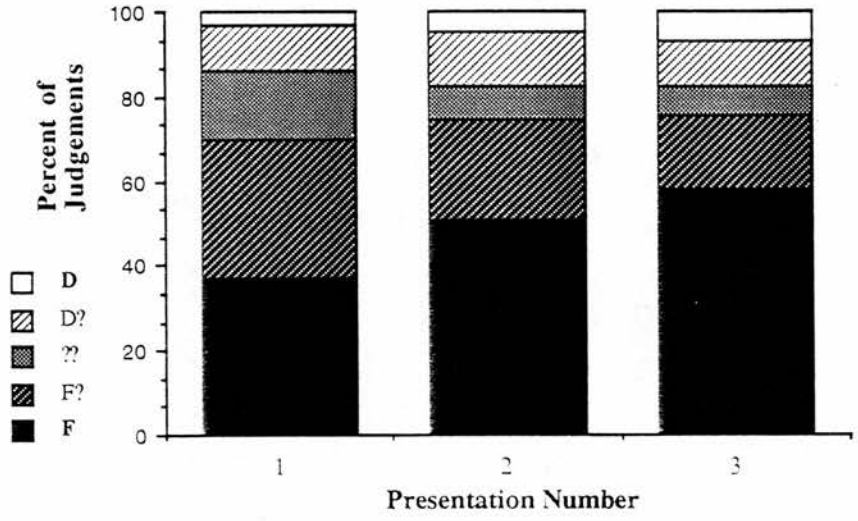

Figure 6.2. Experiment 4: Distribution of fluency judgements by presentation: fluent stimuli.

Variable	$\bar{X}$	sd_s	sd_m
D1	3.14	0.437	0.796
D2	3.30	0.459	0.943
D3	3.36	0.474	0.993
F1	2.10	0.445	0.565
F2	1.98	0.510	0.671
F3	1.90	0.498	0.692

Table 6.3. Experiment 4: Cell means and standard deviations for 2-way ANOVA (fluency by presentation). D = "Disfluent", F = "Fluent" and 1,2,3 are first, second and third presentations.

Two-way analyses of variance with repeated measures for presentation (first, second, third) and fluency (disfluent, fluent) conditions, with cells as totals of disfluency judgements confirmed the above observations. A highly significant main effect of **fluency** was found both by subjects and by materials ($F_{1(1,11)} = 101.11$, $p < 0.0001$; $F_{2(1,29)} = 42.81$, $p < 0.0001$; $MinF'_{(1,39)} = 30.076$, $p < 0.01$). Since in fluent cases, the mean disfluency judgement fell on later presentations and in disfluent cases the mean rose, no main effect of **presentation** was found but there was a highly significant interaction of **fluency by presentation** ($F_{1(2,22)} = 37.73$, $p < 0.0001$; $F_{2(2,58)} = 15.33$, $p < 0.0001$; $MinF'_{(2,80)} =, 10.90$ $p < 0.01$). (Means in table 6.3, illustrated in figure 6.3.)

Post hoc (Scheffé) tests for the **fluency** effect showed that differences between fluent and disfluent stimuli on all three presentations were significant at $p < 0.05$. In addition, Sign tests on the cell means showed that for all subjects the mean disfluency judgement for each presentation was greater for disfluent stimuli than for fluent stimuli and that for the first presentation, means for 27 of the 30 disfluent stimuli, and for the subsequent presentations, means for 28 of the disfluent stimuli were greater than for their fluent matched pairs.

The results clearly support the experimental hypothesis. Subjects were able to distinguish disfluent from fluent stimuli on the basis of the prosodic information in the continuation, given all linguistic information up to the point of interruption.

Cues available to the listener in the prosody of the disfluent stimuli should

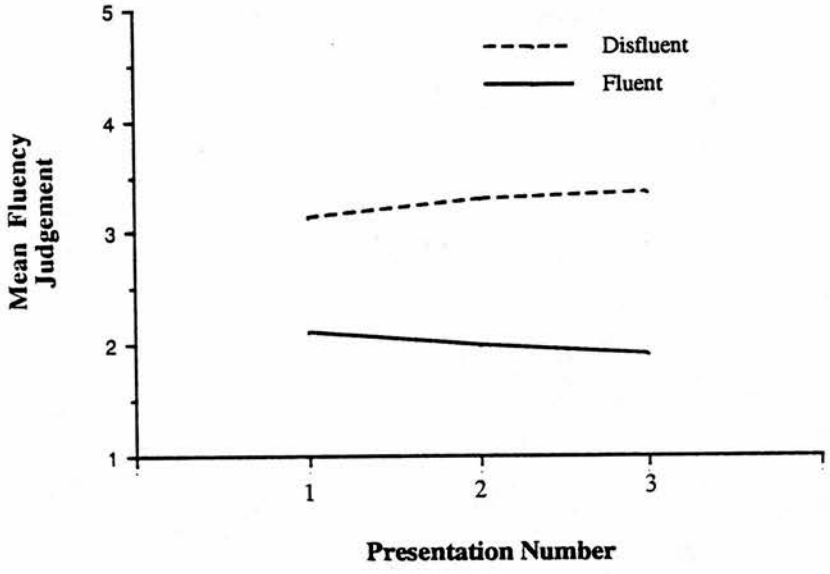

Figure 6.3. Experiment 4: Cell means of fluency judgments for fluent vs. disfluent stimuli by presentation number.

include the timing and intonation of the utterances. If the interruption contains an unexpected pause, which, as we have seen already in the previous experiments, was present in several of the stimuli, it may be that subjects perceived the pause as a sign of disfluency. A discontinuous intonation curve may also have been perceived and resulted in more judgements of "disfluent". In order to find out if such cues had any effect on disfluency judgements, we looked for correlations between judgement totals and pause length and between judgement totals and differences in F_0 before and after the interruption point. Pause length was measured in milliseconds from the offset of the word before the pause to the onset of the following word, allowing 50msec for word initial stop consonants where they occurred. Two sets of F_0 measurements were used: the first was the difference in Hz between the last voiced section of speech before the interruption and the first voiced section after the interruption; the second was the difference in Hz between the last F_0 peak before the interruption and the first peak after the interruption. An additional analysis looked for a correlation between the length in syllables of the reparandum and the disfluency judgement total: it was possible that, with a longer reparandum, the difference in prosody between the end of the reparandum and the onset of the continuation may be more striking than with a shorter reparandum for a combination of reasons. Reparandum length was measured in syllables and in 3 cases half-syllables were counted where the last word of the reparandum appeared to be cut off before an anticipated syllable was completed.

Longer pauses produced higher judgement totals ($r = 0.507$, $N = 30$, $p = 0.004$). No correlation was found between judgement totals and either F_0 measurement. Longer reparanda also produced higher judgement totals ($r = 0.382$, $N = 30$, $p = 0.037$).

In the absence of pause or in the presence of a pause of less than 100msec, subjects were still able to judge disfluency correctly. The mean disfluency judgement for stimuli with a pause was 3.76, at the third presentation, and without a pause, 3.09, compared with a means of 1.90 for fluent stimuli (table 6.4). Analyses of variance excluding the 12 stimuli which contained pauses of greater than 100msec at the interruption point showed strong main effects of **fluency** ($F_{1(1,11)} = 45.75$, $p < 0.0001$; $F_{2(1,17)} = 12.81$, $p = 0.0023$; $MinF'_{(1,24)} = 10.01$, $p < 0.01$) and significant interactions of **fluency by**

Pause	Variable	X	sd_s	sd_m
Present	D1	3.46	0.528	0.764
($n = 12$)	D2	3.71	0.611	0.825
	D3	3.75	0.584	0.882
	F1	1.95	0.493	0.419
	F2	1.82	0.553	0.488
	F3	1.77	0.500	0.541
Absent	D1	2.93	0.474	0.765
($n = 18$)	D2	3.02	0.496	0.937
	D3	3.09	0.557	0.995
	F1	2.20	0.473	0.635
	F2	2.08	0.517	0.764
	F3	1.98	0.534	0.779

Table 6.4. Experiment 4: Means and standard deviations of disfluency judgements for disfluent stimuli with and without pause at interruption and matched sets of fluent stimuli (not containing pauses): D = "Disfluent", F = "Fluent" and 1, 2, 3 are first, second and third presentations.

presentation ($F_{1(2,22)} = 31.21$, $p < 0.0001$; $F_{2(2,34)} = 7.19$, $p = 0.0025$; $MinF'_{(2,47)} = 5.844$, $p < 0.01$).

Post Hoc (Scheffé) tests suggested that the differences between fluent-disfluent pairs for the second and third presentations had the greatest effect on the *fluency* main effect. Only the difference between means of the fluent-disfluent pairs for the first presentation were not greater than t'_{crit} for $p < 0.05$.

6.2.3 Discussion

Disfluent stimuli were presented with the signal low-pass filtered from the point of interruption to the end, so that only prosodic information was audible for the latter part of the stimuli. Matched fluent stimuli were treated in the same way, with low-pass filtering from a point equivalent to the interruption point. Subjects were asked to make a judgement on a scale of 1 to 5 as to the fluency of the stimuli at the point at which the filter was applied. The experimental hypothesis was that subjects would be able to distinguish fluent from disfluent stimuli on the basis of the prosody of the continuation and its relation to the original utterance.

The results support the experimental hypothesis. Mean disfluency judgements (on third presentation) of 1.90 for fluent stimuli and 3.36 for disfluent stimuli show that responses differed significantly between fluency conditions.

The effect of the presence of pause at the interruption point was investigated. It was found that subjects were more secure in judging sentences to be disfluent when a pause accompanied the interruption, but that in the absence of a pause, subjects still perceived disfluency correctly. Longer reparanda also produced higher disfluency judgements, but no direct correlation was found between the disfluency judgements and the difference in F_0 values before and after the interruption.

This result supports the finding from Experiment Three, that it is possible to detect disfluency without having accessed lexical and syntactic information in the speech signal following the interruption. The results of these experiments constitute empirical evidence to support the suggestion that prosodic information has an important rôle in the understanding of disfluent speech. The question remains as to when this information is used. However, Experiment Three suggests that

disfluency can be perceived before the end of the first word of the continuation. In the current experiment, however, subjects heard low-pass filtered speech over a longer stretch of speech after the interruption, so that their judgements may not have been made on the basis of the early post-interruption information used by subjects in Experiment 3. A further experiment was therefore designed, to assess the availability of prosodic information early in the continuation.

6.3 Experiment 5 - 35ms gating with low-pass filtered speech

6.3.1 Introduction

In Experiment 3, the two words on either side of the disfluent interruption were presented incrementally with 35ms gates. It was found that listeners were often able to detect disfluency before they had recognised the word after the interruption. In Experiment 4, prosodic information present in low-pass filtered speech was shown to be sufficient for listeners to judge the fluency of an utterance. The experiment described here combines both of these methods of presentation, 35msec gating and low-pass filtered speech, in order to investigate whether the prosodic information present in low-pass filtered speech is sufficient to allow detection of disfluency before the end of the first word of the continuation.

The experimental hypothesis is that listeners can detect disfluency within the first word of the continuation even when low-pass filtering has removed all but prosodic information from the two words immediately surrounding the interruption.

6.3.2 Method

Materials

Materials were the same utterances as those used in the previous experiment with low-pass filtered speech: 30 disfluent utterances and 30 spontaneous fluent utterances taken from a corpus of 6 spontaneous conversations, digitally recorded in a studio. The disfluent utterances had been selected as a representative sample of the distribution of disfluencies in the corpus as a whole. The fluent utterances had been selected to match the disfluent ones for structure and prosody as far as possible.

Design

Because of the long running-time for this type of experiment, the materials were prepared for presentation to 4 subject groups.

The organisation of the materials into the four groups was decided on the following basis: of the 30 disfluent-fluent pairs, 2 had been excluded from the previous 35msec gating experiment, but were included in this one; these two pairs were assigned to all four subject groups; the remaining 28 pairs were assigned evenly, using a latin square, to the four groups. Thus, each subject group was presented with both members of a total of 9 pairs, consisting of 2 from each of 3 more speakers and 1 from each of 3 speakers. The materials were blocked by speaker. The order of presentation with respect to fluency was random.

Material Preparation

The materials were sampled at 16kHz and prepared for the experiment using a Sun Sparcstation and ESPS/Xwaves+ software.

The crucial words in this experiment, as in the previous 35msec gating experiment, were the two words on either side of the interruption point in disfluent utterances and the equivalent point in the fluent controls: it was these words that were low-pass filtered and presented in gated form.

The utterances were low-pass filtered from the onset of the word preceding the interruption point in the disfluent utterances and from an equivalent point in their fluent pairs. The filtering method was the same as that used in Experiment 4 (see page 163).

The filtered sections were *gated* at intervals of 35msec from the same point and until the end of the word following the interruption. Speech after the end of this word was not presented in this experiment. The gating was performed by means of a simple computer programme, which was used to output the speech to Betamax video tapes, with 445Hz tones preceding each new presentation of a stimulus, allowing adequate time between presentations for subjects to make their judgements. The gating programme avoided the problem of sound distortion at the end-points of each gated presentation by smoothly decreasing the intensity to zero over the last 1.5msec.

Four separate tapes were made, one for each subject group.

Task

At each gated presentation, subjects were asked to give a judgement as to the fluency of the utterance at that point. The judgement was to be given by placing a cross in a circle on a five-point continuum between "fluent" and "disfluent".

A separate answer sheet was provided for each item. Each answer sheet was headed with the transcript of the piece of conversation that provided the context for the utterance being tested, followed by a blank line and the beginning of the test utterance. Responses were marked on a matrix made up of rows of five circles, each row being bounded by "FLU" and "DIS". The matrix was the same size for each item, consisting of fifty rows, which provided more than enough space for any one item and meant that subjects had no idea how many presentations of any item they were to expect.

Subjects

Subjects were 41 members of the students and staff of the university of Edinburgh, making up four groups of 10 and one of 11. All were native speakers of English and none reported having hearing disorders.

Procedure

Subjects were seated in individual listening booths in the department of linguistics. Up to four subjects were tested at a time. The booths were equipped with high-quality headphones, answersheets and pens. Full instructions were given orally, from a script, after which subjects were allowed to ask questions about their task. A practice test, using speech from a speaker not included in the actual experiment, was performed and discussed before the actual test began.

Before items for a new speaker were presented, subjects heard about 10-15 seconds of speech by that speaker, in order to familiarise them with the voice. Each item was announced on the tape by its number and each gated presentation was preceded by a warning tone.

The experiment lasted approximately one hour altogether. The tape was paused for a few minutes after half an hour to give subjects a short rest.

6.3.3 Results

A brief overview of the overall results is given first. This is followed by analysis and comparison of judgements at crucial points in the materials by means of first non-parametric and then parametric statistical tests.

The analysis concentrates on comparing the disfluency judgements at crucial points in the disfluent stimuli with control points in the fluent stimuli. Under the null hypothesis, we expect to find no difference in disfluency judgements at these points. If the experimental hypothesis is to be accepted, there must be a tendency for subjects to give more "fluent" judgements at crucial points in the fluent utterances than in the disfluent utterances and more "disfluent" judgements in the disfluent utterances.

Overview

A total of 12,672 disfluency judgements was obtained from 41 subjects who were each presented with 18 experimental items with between 7 and 42 gated presentations (mean = 17.17 gates per item).

In order to facilitate direct comparisons between fluent and disfluent utterances and between points prior to and following interruptions in the disfluent utterances and equivalent points in the fluent controls, the analysis focused on windows of seven gates in each item. These windows were selected such that the middle (fourth) gate was that which contained the onset of voicing in the speech following the interruption or the equivalent point in the fluent control. So the total number of judgements available in the analysis was 5166 (7 x 18 x 41), half of which were in disfluent utterances and half in fluent.

For the purposes of discussing the results, we will refer to the disfluency judgements as being on a 1-5 scale (where "1" signifies "fluent" and "5", "disfluent"), although they were not specifically numbered on the answer sheets.

The distribution of 1-5 judgements by utterance type gives a first indication of the overall results (Table 6.5). There are more "fluent" judgements in the fluent

Judgement (1-5)	Disfluent		Fluent		All	
	Total	Percent	Total	Percent	Total	Percent
1	745	28.84	1019	39.45	1764	34.14
2	609	23.58	585	22.65	1194	23.11
3	453	17.54	433	16.76	886	17.15
4	324	12.54	269	10.41	593	11.47
5	452	17.50	277	10.72	729	14.11
ALL	2583	100	2583	100	5166	100

Table 6.5. Experiment 5: Distribution of disfluency judgements in the vicinity of the interruption by utterance-type.

utterances and more "disfluent" judgements in the disfluent utterances over the whole window. The difference between distributions of judgements for the two types is significant (χ^2=90.60, df=4, p<0.001).

Non-parametric tests

Since the data consisted of judgements on a 1-5 scale, which we will assume subjects treated as ordinal, we begin the analysis with appropriate non-parametric tests.

Overall effects. The two questions of interest in the initial analysis are whether there was a difference in responses between gates prior to and following the onset of the *continuation* (**place**) and whether there was a difference between responses in disfluent and fluent utterances (**fluency**). The design of the experiment was such that the same subjects responded to both members of the fluent-disfluent pairs in their subset of the data.

To establish whether there were any effects of place and fluency on disfluency judgements overall, by-materials *Friedman* tests were performed, with three conditions for place (disfluency judgements at positions 1,4 and 7 in the selected windows, where position 4 was the gate which included onset of voicing after the interruption) and two conditions for fluency (fluent and disfluent), giving a total of six conditions. The difference in rank totals for the six conditions was found

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	2.4146	1.392	1	5	1263.0	369
D2	2.5907	1.425	1	5	1341.0	369
D3	2.9648	1.500	1	5	1514.0	369
F1	2.1816	1.327	1	5	1149.0	369
F2	2.3279	1.334	1	5	1218.5	369
F3	2.4499	1.462	1	5	1263.5	369

Table 6.6. Experiment 5: Means and rank sums for Friedman test.

to be significant (Xr^2=61.31, p<0.0001, df=5). It is clear from the rank sums displayed in table 6.6 (and from the means illustrated in figure 6.4) that **place** has an effect, with judgements moving towards "disfluent", both in disfluent *and* in fluent utterances. There also appears to be an effect of **fluency**, with higher scores for disfluent utterances. But these higher scores occur for each place, including the gate before the repair onset, so it is unclear from this test whether the trend toward judgements of disfluency over gates between sentence types is greater in the disfluent cases. We will address this question later, with parametric analyses of variance.

The Friedman test shows that there are significant differences somewhere among the six conditions, but does not give a clear picture of which pairs of differences are significant. To show which of the six conditions produced significantly different responses, it is necessary to compare pairs of conditions individually. *Wilcoxon signed-rank* tests were performed on each possible pair of the six conditions[1]. Significant differences (p<0.05) were found between all pairs except for D1:F2, D1:F3, D2:F3 and F2:F3 (Table 6.7).

These results show that subjects were less likely to give low disfluency judgements when they heard the onset of a new word, whether it was the continuation of a fluent utterance or the onset of the fluent continuation in a disfluent utterance, than before this onset. So, there is a **place** effect in both utterance types. In fluent utterances, though, the place effect is restricted to the difference

[1]D1, D2, D3, F1, F2, F3, where D="Disfluent", F="Fluent" and "1", "2" and "3" are gates 3 before, at and 3 after the onset of the repair

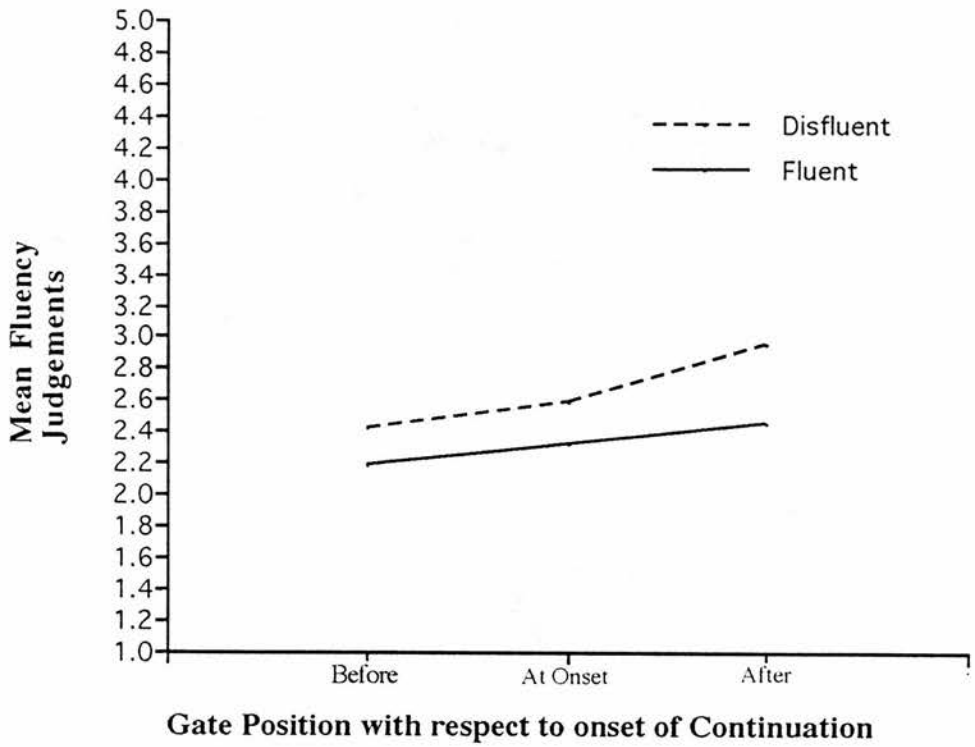

Figure 6.4. Experiment 5: Mean disfluency judgements at three crucial gates.

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	0.0048	1.0000				
D3	0.0000	0.0000	1.0000			
F1	0.0125	0.0000	0.0000	1.0000		
F2	0.3138	0.0046	0.0000	0.0093	1.0000	
F3	0.6772	0.1723	0.0000	0.0021	0.0613	1.0000

Table 6.7. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation.

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	2.6016	1.436	1	5	443.0	123
D2	2.7642	1.510	1	5	463.5	123
D3	3.3089	1.510	1	5	554.0	123
F1	1.8130	1.118	1	5	327.5	123
F2	2.1626	1.276	1	5	379.0	123
F3	2.4390	1.449	1	5	416.0	123

Table 6.8. Experiment 5: Means and rank sums for Friedman test - With-pause condition.

between the pre-onset judgement and the at-onset and post-onset judgements, and is not significant between at-onset and post-onset, whereas, in the disfluent utterances, the place effect occurs between all three positions.

The results also show an effect of **fluency** at all three positions. The experimental hypothesis predicts a difference in disfluency judgements at and after the onset of the continuation, but not, as appears here, prior to this onset ($\overline{X}_{D1} > \overline{X}_{F1}$). We explore a possible reason for this result in the next section: the presence of a silent pause in some of the stimuli may have affected the disfluency judgements at the initial point in the analysis window.

The pause effect. Some of the disfluent utterances contained silent pauses before the onset of the continuation. It is possible that subjects were able to perceive the pause and that this affected their disfluency judgements at (and before) the beginning of the analysis window. To test this hypothesis, the Friedman and Wilcoxon tests were run again, separating pairs of stimuli whose disfluent member contained silence of greater than 100ms at the interruption (N=12) from those which contained less or no silence (N=18).

In the with-pause condition, the difference in rank sums in the six conditions was, as expected, found to be highly significant (Xr^2=69.61, p<0.0001, df=5) (Table 6.8). In the Wilcoxon tests, the disparity between disfluent and fluent judgements was larger than in the mixed condition: only the D1:D2, D1:F3 and D2:F3 pairs do not differ significantly at p<0.05 (two-tailed) (Table 6.9).

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	0.0805	1.0000				
D3	0.0000	0.0000	1.0000			
F1	0.0000	0.0000	0.0000	1.0000		
F2	0.0058	0.0006	0.0000	0.0000	1.0000	
F3	0.2995	0.0626	0.0000	0.0001	0.0124	1.0000

Table 6.9. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation - With-pause condition.

The pause-effect hypothesis is supported by the results in the no-pause condition: the Friedman test still shows a significant rank-sum ordering, but at a lower level of significance than in the other conditions (Xr^2=69.61, p<0.0001, df=5) (Table 6.10). The Wilcoxon tests show that the fluency effect in the mixed condition was mostly caused by the presence of pause prior to the analysis window: the same-place fluent-disfluent pairs are not significantly different except in the case of the after-onset place (D3:F3) (Table 6.11). In addition, the place effect found in the first two sets of Wilcoxon tests is only present in the disfluent utterances in this set: D3 is given higher judgements than D2 and D1, whereas F3 does not differ from F1 and F2.

So, from these tests, we can conclude that subjects were influenced in their judgements by the presence of silent pauses at the interruption point, which results in higher judgements earlier in the analysis window: where there was no pause, the judgements at the start of the analysis window in disfluent stimuli did not differ from those in fluent stimuli.

Prior context effect. It is possible that information in words before the onset of the low-pass filtered section of speech may have given subjects a cue to the presence of disfluency. Although evidence from the no-pause condition Wilcoxon tests, above, suggests otherwise, since there is no difference in disfluency judgements between disfluent and fluent utterances at the beginning of the analysis window, it was still not possible to rule out the possibility. If it were the case

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	2.3211	1.363	1	5	820.0	246
D2	2.5040	1.375	1	5	877.5	246
D3	2.7927	1.469	1	5	960.0	246
F1	2.3658	1.387	1	5	821.5	246
F2	2.4105	1.358	1	5	839.5	246
F3	2.4553	1.472	1	5	847.5	246

Table 6.10. Experiment 5: Means and rank sums for Friedman test - no-pause condition.

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	0.0241	1.0000				
D3	0.0000	0.0002	1.0000			
F1	0.6422	0.2191	0.0005	1.0000		
F2	0.4180	0.3570	0.0009	0.5242	1.0000	
F3	0.2269	0.7409	0.0040	0.4801	0.7001	1.0000

Table 6.11. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation - No-pause condition.

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	1.8753	1.154	1	5	1246.0	369
D2	1.9675	1.222	1	5	1302.5	369
D3	2.0732	1.282	1	5	1364.5	369
F1	1.8807	1.205	1	5	1235.0	369
F2	1.9350	1.234	1	5	1265.0	369
F3	2.0217	1.235	1	5	1336.0	369

Table 6.12. Experiment 5: Data from Friedman test: fluency by place for first three gates – all data.

that prior context affected the disfluency judgements, we would expect there to be higher scores at an early point in the presentation of a stimulus. So in order to test this hypothesis, judgements for the very first three gates of all stimuli were compared, by means of first Friedman tests and then Wilcoxon tests.

Friedman tests were performed on all stimuli and separately on the with-pause and no-pause data. In all three cases, no overall difference was found between the six conditions (All data: Xr^2=10.37, p=0.0654; with-pause: Xr^2=1.24, p=0.9413; no-pause; Xr^2=10.23, p=0.069; df=5), although the means and rank sums suggest that the place effect is present even at the beginning of the stimuli (Tables 6.12, 6.13 and 6.14). To support the hypothesis, we would expect the place effect to combine with generally higher judgements in the disfluent condition, giving significant differences in the rank sums. So the Friedman tests do not support the prior-context-effect hypothesis.

Wilcoxon tests also do not allow us to reject the null hypothesis. No differences are found in any of the 3 sets of Wilcoxon tests between fluent-disfluent same-place pairs. The only significant differences found ($p<0.05$) are between different place variables. This confirms the presence of a place effect as observed in the Friedman tests (Tables 6.15, 6.16 and 6.17)

In conclusion, no effect of prior context on disfluency judgements was observed: there was no significant difference between disfluency judgements in fluent and disfluent utterances for the first three presentations of the stimuli.

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	1.8618	1.126	1	5	425.5	123
D2	1.8943	1.151	1	5	434.5	123
D3	1.9593	1.176	1	5	442.5	123
F1	1.8536	1.192	1	5	418.0	123
F2	1.9105	1.274	1	5	421.5	123
F3	1.9756	1.251	1	5	441.0	123

Table 6.13. Experiment 5: Data from Friedman test: fluency by place for first three gates – with-pause condition.

Variable	Mean	S.D.	Min	Max	Rank Sum	N
D1	1.8821	1.170	1	5	820.5	246
D2	2.0041	1.257	1	5	868.0	246
D3	2.1301	1.331	1	5	922.0	246
F1	1.8943	1.214	1	5	817.0	246
F2	1.9471	1.216	1	5	843.5	246
F3	2.0447	1.230	1	5	895.0	246

Table 6.14. Experiment 5: Data from Friedman test: fluency by place for first three gates – no-pause condition.

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	**0.0087**	1.0000				
D3	**0.0000**	**0.0015**	1.0000			
F1	0.9335	0.1567	**0.0075**	1.0000		
F2	0.3231	0.5576	0.0597	0.0757	1.0000	
F3	**0.0376**	0.4640	0.4631	**0.0016**	**0.0098**	1.0000

Table 6.15. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation for the first three gates – all data.

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	0.7172	1.0000				
D3	0.3147	0.3318	1.0000			
F1	0.8709	0.6725	0.3745	1.0000		
F2	0.7884	0.9269	0.7166	0.3152	1.0000	
F3	0.4932	0.6836	0.9656	0.1554	0.3914	1.0000

Table 6.16. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation for the first three gates – with-pause condition.

	D1	D2	D3	F1	F2	F3
D1	1.0000					
D2	**0.0038**	1.0000				
D3	**0.0000**	**0.0024**	1.0000			
F1	0.7877	0.1408	**0.0075**	1.0000		
F2	0.2399	0.4341	**0.0342**	0.1475	1.0000	
F3	**0.0305**	0.5061	0.3838	**0.0043**	**0.0101**	1.0000

Table 6.17. Experiment 5: Two-tail level of significance of Wilcoxon signed rank tests using the normal approximation for the first three gates – no-pause condition.

Non-parametric tests: Conclusion. From these non-parametric tests we can conclude that within the 7-gate analysis window the following effects can be observed: a **place** effect in disfluent utterances, by which the disfluency judgements are higher on the scale the later they occur in the window; a similar but slightly less stable **place** effect in fluent utterances; a **fluency** effect at all points in the window, which is apparently due to the presence of pausing at the disfluent interruption, since it is not found in the first position in the window where only stimuli with no silent pause are tested; a **fluency** effect in the last gate of the analysis window, which is present even in the set of utterances with no pause at the interruption. It was also established that prior context had no immediate effect on judgements of fluency.

Parametric Tests

To examine interactions between fluency and place (and subsequently other parameters) it was necessary to treat the 1-5 judgements as being on an interval scale.

The cells for these tests were made up of the mean of the subjects' 1 to 5 judgements for each gate within each stimulus. As in the previous tests, cells were made up of judgements at three places in the seven-gate analysis window, (first, middle and last), and the two fluency conditions.

Overall effects

Two-way analyses of variance with repeated measures for fluency and place by subjects and by materials (the parametric equivalent of the Friedman test, page 179) confirmed the **fluency** effect noted in the non-parametric tests: disfluent utterances received higher mean disfluency judgements than fluent utterances ($F_{1(1,40)} = 11.04$, $p = 0.0019$; $F_{2(1,29)} = 9.02$, $p = 0.0055$). The **place** effect was also confirmed: mean disfluency judgements rose throughout the analysis window ($F_{1(2,80)} = 15.66$, $p < 0.0001$; $F_{2(2,58)} = 26.38$, $p < 0.0001$). The interaction of **fluency by place** was also significant, showing that the increase in the disfluency judgements observed for both utterance types was significantly different, and, according to the means, higher in disfluent utterances than in

Variable	Mean	S.D.	Min	Max	N
D1	2.3383	0.575	1.50	3.60	30
D2	2.5203	0.599	1.36	3.64	30
D3	2.9619	0.582	2.20	4.40	30
F1	2.0251	0.657	1.30	3.82	30
F2	2.2242	0.593	1.40	3.70	30
F3	2.3852	0.661	1.20	4.20	30

Table 6.18. Experiment 5: Cell means and standard deviations for ANOVA by materials.

Variable	Mean	S.D.	Min	Max	N
D1	2.4146	0.696	1.11	3.89	41
D2	2.5908	0.755	1.22	3.89	41
D3	2.9648	0.876	1.22	4.56	41
F1	2.1816	0.623	1.00	3.89	41
F2	2.3279	0.670	1.11	3.67	41
F3	2.4499	0.746	1.22	4.00	41

Table 6.19. Experiment 5: Cell means and standard deviations for ANOVA by subjects.

fluent utterances ($F_{1(2,80)} = 7.56$, $p = 0.001$; $F_{2(2,58)} = 4.04$, $p = 0.0228$) (see tables 6.18 and 6.19).

The pause effect. To determine whether the difference in the rise in disfluency judgement means over the analysis window remains significant whether or not a silent pause is present in the disfluent stimulus, ANOVAs were performed for the separate with-pause and no-pause conditions.

In the with-pause condition significant **fluency** and **place** effects were once again found by subjects and by materials (**Fluency**: $F_{1(1,40)} = 23.74$, $p < 0.0001$; $F_{2(1,11)} = 15.48$, $p = .0024$; **Place**: $F_{1(2,80)} = 21.39$, $p < 0.0001$; $F_{2(2,22)} = 15.94$, $p = 0.0001$) but no interaction effect ($F_{1(2,80)} = 1.25$, $p = 0.2912$; $F_{2(2,22)} = 1.14$, $p = 0.3392$) (Tables 6.20 and 6.21).

In the no-pause condition, the results are different: the removal of utterances

Variable	Mean	S.D.	Min	Max	N
D1	2.5925	0.501	2.00	3.60	12
D2	2.7533	0.497	2.10	3.64	12
D3	3.3050	0.492	2.60	4.40	12
F1	1.8183	0.437	1.40	2.70	12
F2	2.1675	0.651	1.40	3.70	12
F3	2.4350	0.838	1.20	4.20	12

Table 6.20. Experiment 5: Cell means and standard deviations for ANOVA by materials in with-pause condition.

Variable	Mean	S.D.	Min	Max	N
D1	2.5833	0.956	1.00	4.67	41
D2	2.7520	1.069	1.00	5.00	41
D3	3.2480	1.150	1.00	5.00	41
F1	1.7602	0.722	1.00	3.75	41
F2	2.0630	0.933	1.00	4.25	41
F3	2.3211	0.998	1.00	4.50	41

Table 6.21. Experiment 5: Cell means and standard deviations for ANOVA by subjects in with-pause condition.

Variable	Mean	S.D.	Min	Max	N
D1	2.1689	0.571	1.50	3.44	18
D2	2.3650	0.624	1.40	3.55	18
D3	2.7332	0.530	2.00	4.09	18
F1	2.1629	0.750	1.30	3.82	18
F2	2.2621	0.568	1.60	3.64	18
F3	2.3519	0.539	1.50	3.27	18

Table 6.22. Experiment 5: Cell means and standard deviations for ANOVA by materials in no-pause condition.

Variable	Mean	S.D.	Min	Max	N
D1	2.3453	0.727	1.00	4.00	41
D2	2.5221	0.759	1.29	4.33	41
D3	2.8085	0.892	1.33	5.00	41
F1	2.3966	0.727	1.00	4.00	41
F2	2.4341	0.678	1.17	4.00	41
F3	2.4606	0.812	1.17	4.33	41

Table 6.23. Experiment 5: Cell means and standard deviations for ANOVA by subjects in no-pause condition.

with pauses from the analysis results in there being no significant overall difference in judgements for fluent versus disfluent utterances in either by-subjects or by-materials analyses ($F_{1(1,40)} = 1.80$, $p = 0.1868$; $F_{2(1,17)} = 1.03, p = 0.325$). However, the **place** effect is still significant ($F_{1(2,80)} = 5.22$, $p = 0.0074$; $F_{2(2,34)} = 11.67$, $p = 0.0001$ as is the **fluency by place** interaction ($F_{1(2,80)} = 8.76$, $p = 0.0004$; $F_{2(2,34)} = 4.28$, $p = 0.022$) (Tables 6.22 and 6.23). This last result is interesting as it shows that, although the fluency effect is not significant, the *rise* in mean disfluency judgement scores is greater in the case of disfluent utterances than in fluent utterances, thus supporting the experimental hypothesis.

The Wilcoxon tests on page 182 showed that the only significant difference for fluency in the no-pause condition was found at the third place in the analysis window. To determine whether a similar result obtained in a parametric test,

Place	$\overline{X}_D$	$\overline{X}_F$	t	α	df
By Materials					
1	2.1689	2.1629	0.03	0.9746	17
2	2.3650	2.2621	0.56	0.5808	17
3	2.7332	2.3519	2.28	0.0360	17
By Subjects					
1	2.3453	2.3966	-0.58	0.5635	40
2	2.5221	2.4341	0.81	0.4241	40
3	2.8085	2.4606	2.66	0.0111	40

Table 6.24. Experiment 5: t test by materials comparing judgements in fluent versus disfluent utterances with no pause at interruption point.

using the cell means used in the ANOVAs, related t tests were performed, comparing judgements for the disfluent and fluent utterances at each of the three points in the analysis window. The results support those found in the Wilcoxon tests: there was no significant difference between mean scores for the two conditions in the first or the second places (Place 1 (by materials): t=0.03, df=17, p=0.9746,; Place 1 (by subjects): t=−0.58, df=40, p=0.5635; Place 2 (by materials): t=0.56, df=17, p=0.5808; Place 2 (by subjects): t=0.81, df=40, p=0.4241); the difference between mean scores for the third position, three gates after the onset of the continuation in the disfluent utterances, was significant, the mean scores in the disfluent utterances being higher than those in their fluent controls (by materials: t=2.28, df=17, p=0.0360; by subjects: t=2.66, df=40, p=0.0111) (Table 6.24).

Finally, a direct comparison between disfluency judgements on disfluent utterances with pauses and those without was made, by means of a two-way ANOVA by subjects, with two pause conditions and three place conditions. The mean disfluency judgement was found to be higher in the with-pause condition in all three places, making the pause effect significant (F=5.41, p=0.0252, df=1,40); the place effect was highly significant (F=25.68, p<0.0001, df=2,80); the interaction of pause and place was not found to be significant – there was no significant difference between the rises in disfluency judgements between the two pause conditions ($F_{(2,80)} = 1.68$, $p = 0.1936$) (Table 6.25).

Variable	Mean	S.D.	Min	Max	N
D1	2.5833	0.955	1.00	4.67	41
D2	2.7520	1.069	1.00	5.00	41
D3	3.2479	1.151	1.00	5.00	41
F1	2.3453	0.727	1.00	4.00	41
F2	2.5221	0.759	1.28	4.33	41
F3	2.8085	0.892	1.33	5.00	41

Table 6.25. Experiment 5: Cell means and standard deviations for ANOVA by subjects in with-pause condition.

The differences between the results in the with-pause and no-pause conditions and the differences found in the direct comparison of scores by the same subjects for utterances with pauses versus those without, support the pause-effect hypothesis: the presence of a pause before the onset of the continuation caused subjects to give higher disfluency judgements earlier than in the no-pause condition.

Prior context effect. Friedman and Wilcoxon tests described in the previous section showed no effect of prior context on disfluency judgements. To confirm these results and to test for interactions, the equivalent parametric tests were also performed on the data. The null hypothesis is that there is no effect of prior context on disfluency judgements in the different fluency conditions.

Two way ANOVAs by materials and by subjects were performed, comparing the mean disfluency judgements for each of the first three gates of disfluent utterances with those from fluent utterances. No significant differences were found between judgements for **fluency** ($F_{1(1,40)} = 0.26,\ p = 0.6104$; $F_{2(1,29)} = 0.06, p = 0.8113$). mean judgements increased over **place** in both disfluent and fluent conditions ($F_{1(2,80)} = 8.94,\ p < 0.001$; $F_{2(2,58)} = 12.81,\ p < 0.0001$); there was no difference in the rise in disfluency judgements across place between disfluent and fluent conditions ($F_{1(2,80)} = 0.77,\ p = 0.4664$; $F_{2(2,58)} = 0.12,\ p = 0.8830$).

To determine whether the pause effect was related to a prior context effect, ANOVAs with the same conditions for place and fluency were also performed on the separate no-pause and with-pause sets. If such an effect were present, it

Variable	Mean	S.D.	Min	Max	N
D1	1.8073	0.382	1.50	2.82	30
D2	1.8832	0.477	1.30	2.91	30
D3	1.9749	0.511	1.30	3.20	30
F1	1.7933	0.532	1.10	3.82	30
F2	1.8466	0.531	1.10	3.82	30
F3	1.9388	0.564	1.10	3.73	30

Table 6.26. Experiment 5: Cell means and standard deviations for ANOVA by materials for first three gates.

Variable	Mean	S.D.	Min	Max	N
D1	1.8753	0.881	1.00	3.78	41
D2	1.9675	0.861	1.00	3.78	41
D3	2.0732	0.793	1.00	3.89	41
F1	1.8808	0.811	1.00	3.78	41
F2	1.9350	0.810	1.00	3.67	41
F3	2.0217	0.745	1.00	3.78	41

Table 6.27. Experiment 5: Cell means and standard deviations for ANOVA by subjects for first three gates.

Variable	Mean	S.D.	Min	Max	N
D1	1.8492	0.441	1.40	2.82	12
D2	1.9136	0.543	1.30	2.91	12
D3	1.9893	0.545	1.30	3.20	12
F1	1.5848	0.202	1.10	1.82	12
F2	1.6977	0.257	1.10	2.10	12
F3	1.8129	0.367	1.10	2.40	12

Table 6.28. Experiment 5: Cell means and standard deviations for ANOVA by materials for first three gates – with-pause condition.

Variable	Mean	S.D.	Min	Max	N
D1	1.8753	0.881	1.40	2.56	18
D2	1.9675	0.861	1.40	2.85	18
D3	2.0732	0.793	1.30	3.07	18
F1	1.8808	0.811	1.30	3.82	18
F2	1.9350	0.810	1.30	3.82	18
F3	2.0217	0.745	1.40	3.73	18

Table 6.29. Experiment 5: Cell means and standard deviations for ANOVA by materials for first three gates – no-pause condition.

would be expected to cause higher mean disfluency judgements in the disfluent stimuli in the with-pause set. If no such effect were present, we would expect only the usual place effect to show significance.

The by-materials ANOVAs fail to reject the null hypothesis: there is no significant difference for **fluency** in either the with-pause condition ($F_{2(1,11)} = 1.95$, $p = 0.1899$) or the no-pause condition ($F_{2(1,17)} = 0.34$, $p = 0.5660$); the **place** effect is observed in both conditions (with-pause: $F_{2(2,22)} = 6.15$, $p = 0.0075$; no-pause: $F_{2(2,34)} = 6.43$, $p = 0.0043$); no interaction of **fluency by place** is observed in either condition (with-pause: $F_{2(2,22)} = 0.45, p = 0.6424$; no-pause: $F_{2(2,34)} = 1.38$, $p = 0.2641$) (tables 6.28 and 6.29).

The by-subjects analyses differ. In the no-pause condition, the results are the same as in the by-materials analyses, (Fluency: $F_{1(1,40)} = 2.19, p = 0.1467$; place: $F_{1(2,80)} = 8.06, p < 0.001$; F×P: $F_{1(2,80)} = 3.02, p = 0.0543$) (table 6.31).

Variable	Mean	S.D.	Min	Max	N
D1	1.8455	0.914	1.00	3.33	41
D2	1.9085	0.937	1.00	4.00	41
D3	1.9797	0.868	1.00	4.00	41
F1	1.5549	0.683	1.00	3.00	41
F2	1.6524	0.754	1.00	3.25	41
F3	1.7561	0.707	1.00	3.25	41

Table 6.30. Experiment 5: Cell means and standard deviations for ANOVA by subjects for first three gates – no-pause condition.

Variable	Mean	S.D.	Min	Max	N
D1	1.8830	0.895	1.00	3.67	41
D2	1.9988	0.866	1.00	3.80	41
D3	2.1288	0.803	1.00	3.80	41
F1	2.0483	0.930	1.00	4.17	41
F2	2.0746	0.913	1.00	4.17	41
F3	2.1495	0.847	1.00	4.20	41

Table 6.31. Experiment 5: Cell means and standard deviations for ANOVA by subjects for first three gates – no-pause condition.

In the with-pause condition, the mean disfluency judgements are significantly different between disfluent and fluent utterances ($\overline{X}_D > \overline{X}_F$), producing a fluency effect which, on its own, would support the hypothesis that prior context had an effect on disfluency judgements in utterances that contained pauses (fluency: $F_{1(1,40)} = 9.62$, $p = 0.035$; place: $F_{1(2,80)} = 5.85$, $p = 0.0043$; F×P: $F_{1(2,80)} = 0.51$, $p = 0.6023$) (table 6.30). Examination of the cell means for these tests (table 6.30 and 6.31) suggested that, rather than this effect being caused by higher than expected mean judgements for disfluent utterances, it was caused by *lower* judgements for fluent utterances.

To test for this, T tests were performed comparing the mean judgements by subjects for all fluency and pause conditions at each place. The tests confirmed the observation that disfluency judgements for fluent utterances in the with-pause condition were lower than in any other condition, while those for the disfluent

Place	Comparison	$\overline{X}_1$	$\overline{X}_2$	t	α	N
1	DP-FP	1.8455	1.5549	3.69	0.0007	41
2	DP-FP	1.9085	1.6524	2.70	0.0100	41
3	DP-FP	1.9797	1.7560	2.26	0.0292	41
1	DP-DNP	1.8455	1.8831	-0.60	0.5525	41
2	DP-DNP	1.9085	1.9989	-1.22	0.2298	41
3	DP-DNP	1.9797	2.1288	-1.94	0.0596	41
1	DP-FNP	1.8455	2.0483	-2.48	0.0174	41
2	DP-FNP	1.9085	2.0746	-1.98	0.0542	41
3	DP-FNP	1.9797	2.1495	-2.09	0.0427	41
1	FP-DNP	1.5549	1.8831	-4.14	0.0002	41
2	FP-DNP	1.6524	1.9989	-3.74	0.0006	41
3	FP-DNP	1.7560	2.1288	-3.84	0.0004	41
1	FP-FNP	1.5549	2.0483	-6.17	0.0000	41
2	FP-FNP	1.6524	2.0746	-4.57	0.0000	41
3	FP-FNP	1.7560	2.1495	-4.12	0.0002	41
1	DNP-FNP	1.8831	2.0483	-2.47	0.0179	41
2	DNP-FNP	1.9989	2.0746	-1.14	0.2615	41
3	DNP-FNP	2.1288	2.1495	-0.29	0.7722	41

Table 6.32. Experiment 5: t test by subjects comparing all pairs of judgements for each of first three gates for fluency and presence of pause at interruption point.

utterances in the with-pause condition either did not differ from mean judgements in the no-pause condition or were slightly lower (Table 6.32). It is possible that the results in the with-pause condition are unreliable, because there are so few stimuli per subject (2, 3 or 4 per fluency condition, as opposed to 5, 6 or 7 in the no-pause condition).

No effect of prior context on disfluency judgements in disfluent utterances was observed. A possible effect on utterances which contained a later pause was not confirmed by the non-parametric tests, nor by subsequent analyses, which showed rather that the apparent effect was more likely to be due to unusually low mean scores in the fluent condition in a small part of the data.

Judgement Peaks. The seven gate analysis window was chosen because it allowed comparisons to be made between disfluency judgements for disfluent and fluent utterances at points which were strictly controlled with respect to their distance in time from the onset points of fluent continuations and the equivalent points in fluent utterances. The last point in the analysis window was found to have the highest mean disfluency judgement for both fluency conditions (see, for example, Table 6.18). However, for individual utterances, this point was not necessarily the highest mean judgement within the window nor over the whole utterance.

Within the analysis window, the last gate had the highest mean disfluency judgement score in 20 of the 30 disfluent utterances and in 15 of the fluent utterances. Taking into account all gates within or after the window, the last gate of the window had the highest mean score in 6 disfluent cases and 9 fluent. The highest mean score for disfluent utterances preceded the last gate of the window in 7 cases and followed it in 17. For fluent utterances, the highest mean score preceded the last gate of the window in 7 cases and followed it in 14.

Two sets of comparisons were made to examine the differences between disfluent and fluent utterances using the mean disfluency judgement peaks instead of the mean disfluency judgements at the last point in the analysis window: in the first set, the peak for each disfluent utterance was compared with the mean judgement at the equivalent point in the fluent control, where "equivalence" meant that the judgement was taken from the point in the fluent utterance that was the same number of gates away from the onset of the second gated word as the peak point in the disfluent utterance was from the onset of the continuation; in the second set, the peak for each disfluent utterance was compared to the *peak* in its fluent pair wherever it was. In both sets of tests the place condition was also tested using the first gate of the analysis window for both disfluent and fluent utterances. With these data sets, separate analyses of variance were performed for all utterances, for all utterances containing a pause at the interruption and for those utterances with no such pause.

It was expected, given earlier results, that the place effect would be found in all cases and the fluency effect found in most cases, but less so in no-pause utterances. The main question of interest was whether the interaction would still

be found in these conditions, especially in the latter set of analyses where the judgement in the fluent utterance was also the mean peak.

In the first set of analyses, where the mean judgement for fluent and disfluent utterances were at time-matched gates, the results were as expected: the **place** effect was found in all conditions (all data: $F_{1(1,40)} = 26.63$, $p < 0.0001$; $F_{2(1,29)} = 44.23$, $p < 0.0001$; with pause:$F_{1(1,40)} = 27.58$, $p < 0.0001$; $F_{2(1,11)} = 22.21$, $p = 0.0006$; no pause: $F_{1(1,40)} = 12.32$, $p = 0.0011$; $F_{2(1,17)} = 23.99$, $p = 0.0001$). The **fluency** effect was found, in all conditions except the no-pause condition by materials (all data: $F_{1(1,40)} = 20.90$, $p < 0.0001$; $F_{2(1,29)} = 20.38$, $p < 0.0001$; with pause: $F_{1(1,40)} = 39.29$, $p < 0.0001$; $F_{2(1,11)} = 32.35$, $p = 0.0001$; no pause: $F_{1(1,40)} = 5.09$, $p = 0.0297$; $F_{2(1,17)} = 4.18$, $p = 0.567$). The interaction of **place by fluency**, showing a greater increase in mean judgement over place in disfluent than in fluent utterances, was found to be significant in all conditions other than the with-pause condition by materials but at a lower level of significance in the with-pause condition by subjects (all data: $F_{1(1,40)} = 22.75$, $p < 0.0001$; $F_{2(1,29)} = 13.74$, $p = 0.0009$; with pause: $F_{1(1,40)} = 5.13$, $p = 0.0289$; $F_{2(1,11)} = 2.51$, $p = 0.1415$; no pause: $F_{1(1,40)} = 21.01$, $p < 0.0001$; $F_{2(1,17)} = 12.01$, $p = 0.003$). The means for these data are in tables 6.33 and 6.34.

In the second set of analyses, comparing fluency-judgement peaks in both types of utterance, **place** effects were found in all conditions (all data: $F_{1(1,40)} = 34.28$, $p < 0.0001$; $F_{2(1,29)} = 70.38$, $p < 0.0001$; with pause: $F_{1(1,40)} = 41.81$, $p < 0.0001$; $F_{2(1,11)} = 29.52$, $p = 0.0002$; no pause: $F_{1(1,40)} = 19.76$, $p = 0.0001$; $F_{2(1,17)} = 36.34$, $p = 0.0001$). Fluency effects were found, except in the no-pause condition (all data: $F_{1(1,40)} = 13.02$, $p = 0.0008$; $F_{2(1,29)} = 12.21$, $p = 0.0015$; with pause: $F_{1(1,40)} = 30.47$, $p < 0.0001$; $F_{2(1,11)} = 29.52$, $p = 0.0002$; no pause: $F_{1(1,40)} = 1.19$, $p = 0.2821$; $F_{2(1,17)} = 1.38$, $p = 0.2562$). Means for the four variables (see tables 6.33 and 6.34) show that there is a greater difference over place in disfluent utterances in all conditions. However, this difference only reaches significance in the no-pause condition in the by-materials analyses although it is significant in all conditions of the by-subjects analyses (all data: $F_{1(1,40)} = 4.94$, $p = 0.0320$; $F_{2(1,29)} = 3.93$, $p = 0.0568$; with pause: $F_{1(1,40)} = 5.13$, $p = 0.0289$; $F_{2(1,11)} = 0.02$, $p = 0.94$; no pause: $F_{1(1,40)} = 7.72$, $p = 0.0083$; $F_{2(1,17)} = 6.09$, $p = 0.0245$).

To examine the individual differences between pairs of variables in this study, *t* tests were performed both by materials and subjects and for the three pause conditions (all data, with-pause, no-pause), comparing the following seven pairs: window-initial with peak in disfluent utterances (D1 vs Dpk); window-initial for both utterance types (D1 vs F1); disfluent peak with equivalent point in fluent utterance (Dpk vs Fm); disfluent peak with fluent peak (Dpk vs Fpk); window-initial with point matched to disfluent peak in fluent utterances (F1 vs Fm); window-initial and peak in fluent utterances (F1 vs Fpk); matched point with fluent peak (Fm vs Fpk). The results are displayed in tables 6.33 and 6.34. The only pairs not found to differ significantly are in the no-pause condition, for D1 vs F1 (as observed above, on page 189) and for F1 vs Fm. Importantly, the comparisons of mean judgements for peaks in the disfluent and fluent conditions show that the judgement peaks in disfluent utterances are significantly higher than those in fluent utterances.

In conclusion, an alternative view of the results from that adopted earlier (looking only at judgements at three points within the selected analysis window) was to take peak points for disfluency judgements in disfluent utterances and compare them with equivalent points and peaks in their controls. The tests showed that the highest judgements in disfluent utterances are still significantly higher than in fluent utterances. In disfluent utterances with a pause at the interruption point the level of significance for this distinction is higher than in the no-pause condition.

Variations in subject behaviour

Subjects displayed a range of different behaviours both in their use of the five-point scale and in their ability to perform the task.

Within the 7-gate analysis window, the five-point disfluency judgement scale was fully used by 27 of the 41 subjects (that is, each of the five points was used more than once by each of these subjects in the whole experiment). Ten subjects only used 4 points, 5 of these never using the "5" point (the "certainly disfluent" end of the scale), 3 never using the "3" point and one each avoiding "2" and "4". The judgements for 2 subjects were restricted to "1" to "3" and

Comparison	$\overline{X_1}$	$\overline{X_2}$	t	α	N
All Data:					
D1-Dpk	2.3383	3.1653	-9.56	0.0000	30
D1-F1	2.3383	2.0251	2.15	0.0403	30
Dpk-Fm	3.1653	2.4072	6.40	0.0000	30
Dpk-Fpk	3.1653	2.6611	4.86	0.0000	30
F1-Fm	2.0251	2.4072	-2.89	0.0073	30
F1-Fpk	2.0251	2.6611	-5.35	0.0000	30
Fm-Fpk	2.4072	2.6611	-4.40	0.0001	30
With Pause:					
D1-Dpk	2.5925	3.4783	-6.19	0.0001	12
D1-F1	2.5925	1.8183	4.57	0.0013	12
Dpk-Fm	3.4783	2.4483	5.42	0.0002	12
Dpk-Fpk	3.4783	2.7342	5.38	0.0002	12
F1-Fm	1.8183	2.4483	-2.70	0.0207	12
F1=Fpk	1.8183	2.7342	-4.73	0.0006	12
Fm-Fpk	2.4483	2.7342	-2.97	0.0127	12
No Pause:					
D1-Dpk	2.1689	2.9565	-7.13	0.0000	18
D1-F1	2.1689	2.1629	0.03	0.9746	18
Dpk-Fm	2.9565	2.3797	4.13	0.0007	18
Dpk-Fpk	2.9565	2.6124	2.53	0.0217	18
F1-Fm	2.1629	2.3797	-1.45	0.1642	18
F1-Fpk	2.1629	2.6124	-3.27	0.0037	18
Fm-Fpk	2.3797	2.6124	-3.17	0.0056	18

Table 6.33. Experiment 5: t test by materials comparing all pairs of judgements for mean judgement-peaks.

Comparison	$\overline{X}_1$	$\overline{X}_2$	t	α	N
All Data:					
D1-Dpk	2.4146	3.1463	-6.31	0.0000	41
D1-F1	2.4146	2.1816	2.48	0.0173	41
Dpk-Fm	3.1463	2.4986	5.57	0.0000	41
Dpk-Fpk	3.1463	2.7127	3.90	0.0004	41
F1-Fm	2.1816	2.4986	-3.03	0.0043	41
F1-Fpk	2.1816	2.7127	-4.51	0.0001	41
Fm-Fpk	2.4986	2.7127	-4.19	0.0001	41
With Pause:					
D1-Dpk	2.5833	3.4776	-6.21	0.0000	41
D1-F1	2.5833	1.7602	4.71	0.0000	41
Dpk-Fm	3.4776	2.2276	6.11	0.0000	41
Dpk-Fpk	3.4776	2.6280	4.63	0.0000	41
F1-Fm	1.7602	2.2276	-2.67	0.0109	41
F1-Fpk	1.7602	2.6280	-4.65	0.0000	41
Fm-Fpk	2.2276	2.6280	-3.60	0.0009	41
No Pause:					
D1-Dpk	2.3453	2.9653	-4.70	0.0000	41
D1-F1	2.3453	2.3966	-0.58	0.5635	41
Dpk-Fm	2.9653	2.5337	3.94	0.0003	41
Dpk-Fpk	2.9653	2.7273	2.14	0.0381	41
F1-Fm	2.3966	2.5337	-1.28	0.2077	41
F1-Fpk	2.3966	2.7273	-3.16	0.0030	41
Fm-Fpk	2.5337	2.7273	-4.20	0.0001	41

Table 6.34. Experiment 5: t test by subjects comparing all pairs of judgements for mean judgement peaks.

2 other subjects used only "1" and "5" (with a single exception for one of these subjects). However, this uneven distribution of judgements across the five-point scale would not affect the statistical analyses, since, even where a subject only used 2 or 3 points on the five-point scale, if their judgements changed during the 7 gates analysed, the change would still contribute to the overall results in both the nonparametric and parametric tests.

The difference between subjects in whether or not they performed the task of identifying disfluency could, however, affect the overall results. If a large number of the subjects were unable to perform the task correctly, then the overall results could be weakened.

To discover how successful individual subjects had been in their task, the distribution of judgements across the five-point scale was compared for fluent and disfluent utterances within the seven-gate analysis window. If the task was performed successfully and disfluencies recognised in the latter half of the window, the distribution of judgements would differ between fluent and disfluent stimuli such that fluent stimuli would receive more "fluent" judgements and fewer "disfluent" judgements than disfluent stimuli, as was found to be the case overall (Table 6.5). Chi-squared tests were performed to examine the distribution of judgements between utterance types for each of the 41 subjects.

The tests showed significant χ^2 values for 34 of the 41 subjects ($p<0.05$). In seven of these cases, however, the distribution was not as expected: rather than disfluent stimuli receiving more "disfluent" judgements than fluent stimuli, they received fewer, so that the distribution of judgements between stimulus types was significantly different but did not support the experimental hypothesis. Following these tests, subjects can be seen as falling into three distinct groups: the majority of subjects (27 or 66%) performed the task as predicted, the distribution of their judgements suggesting that they were able to distinguish fluent from disfluent stimuli; 7 subjects (17.07%) were unable to distinguish between fluent and disfluent stimuli – the distribution of their judgements supports the null hypothesis; the remaining 7 subjects produced unexpected responses, which may have reflected a consistent misunderstanding of the instructions: the distribution of their judgements shows a bias towards more "disfluent" judgements in the fluent stimuli than in the disfluent stimuli. The aggregated distributions

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Disfluent	417	390	316	223	355	1701
	24.4%	*22.9%*	*18.6%*	*13.1%*	*20.9%*	*100%*
Fluent	742	408	280	121	150	1701
	43.6%	*24.0%*	*16.5%*	*7.1%*	*8.8%*	*100%*
Marginal Totals	1159	798	596	344	505	3402
	34.1%	*23.4%*	*17.5%*	*10.1%*	*14.8%*	*100%*

Table 6.35. Experiment 5: Aggregated disfluency judgement distribution for "Correct" group.

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Disfluent	133	125	71	62	50	441
	30.1%	*28.3%*	*16.1%*	*14.1%*	*11.3%*	*100%*
Fluent	141	121	62	76	41	441
	31.9%	*27.4%*	*14.1%*	*17.2%*	*9.3%*	*100%*
Marginal Totals	274	246	133	138	91	882
	31.1%	*27.9%*	*15.1%*	*15.6%*	*10.3%*	*100%*

Table 6.36. Experiment 5: Aggregated disfluency judgement distribution for "Non-significant" group.

of judgements by stimulus type for each of these three groups are shown in tables 6.35, 6.36 and 6.37. The χ^2 statistic for each of the groups reflects that of its members: for the "correct" group, $\chi^2 = 207.177, p < 0.001, df = 4$; for the "non-significant" group, $\chi^2 = 3.21804, p > 0.5, df = 4$; for the "incorrect" group, $\chi^2 = 45.3711, p < 0.001$.

6.3.4 Discussion

The experimental hypothesis was that listeners can detect disfluency within the first word of the continuation even when low-pass filtering has removed all but

Fluency Type	Judgements					Marginal Totals
	Fluent	Fluent?	??	Disfluent?	Disfluent	
Disfluent	195	94	66	39	47	441
	44.2%	*21.3%*	*15.0%*	*8.8%*	*10.7%*	*100%*
Fluent	136	56	91	72	86	441
	30.8%	*12.7%*	*20.6%*	*16.3%*	*19.5%*	*100%*
Marginal Totals	331	150	157	111	133	882
	37.5%	*17.0%*	*17.8%*	*12.6%*	*15.1%*	*100%*

Table 6.37. Experiment 5: Aggregated disfluency judgement distribution for "incorrect" group.

prosodic information from the two words immediately surrounding the interruption. This predicted that disfluency judgements in disfluent utterances would be higher after the onset of the word following the disfluent interruption than before this point and also higher than judgements at equivalent points in fluent control utterances.

The clearest effect found in all the data was a **place** effect: in all utterances, irrespective of fluency, there was found to be a general upwards trend in mean disfluency judgements.

The presence of a silent **pause** at the interruption point was found to have an effect on the **fluency** effect and the **interaction** of place and fluency (see below).

A **fluency** effect was observed in the analysis when all utterances were examined together: disfluent utterances had higher mean disfluency judgements throughout the analysis window and in the judgement-peaks comparisons. This effect was found to be mainly due to the presence of silent pauses at the interruption point, which led subjects to give higher judgements earlier than the first point in the analysis window, and later in the window: in the no-pause condition, no fluency effect was found and the mean judgements at the first point in the analysis window were about equal.

For the experimental hypothesis, the most important effect was the interaction of place and fluency. The rise in disfluency judgements over place in the analysis window, which was present in both utterance types, was found to be higher

in disfluent utterances. This effect was found for all data and in the no-pause condition, but not in the with-pause condition. This suggests that, given that the mean judgements for the first three gates of disfluent utterances with subsequent pauses are not higher than for any other set of utterances but that the mean judgements at the beginning of the analysis window for these utterances are higher than for others, there is still a significant rise in judgements in the pause condition but it was earlier than for those with no pause.

The results for disfluency judgements in the analysis window and at judgement peaks were probably not affected by prior context. No differences were found between judgements for fluent and disfluent utterances within the first three gates, nor between utterances with and without subsequent pause. The place effect was found at this stage in the presentation, mean disfluency judgements rising through the first three gates. A weak fluency effect in the with-pause condition was probably caused by unusually low mean judgements in the fluent controls, rather than higher judgements in the disfluent utterances.

Differences in subject behaviour were observed. Some of these indicate that the task of making disfluency judgements from low-pass filtered speech was difficult for a minority of subjects.

The results support the hypothesis: the majority of subjects were able to detect disfluency within the first word of the continuation from the prosodic characteristics of the speech at that point.

Chapter 7

Acoustic Analysis

7.1 Introduction

Our experimental evidence shows that listeners are able to detect disfluency soon after it has occurred and without the benefit of syntactic information. A literal interpretation of Hindle's discrete editing signal was rejected and it was suggested that instead the main cues available to the listener in detecting disfluency lie in the onset of the repair. Some indications of what cues are available were apparent in the results: a silent pause at the interruption point led to more judgements of "disfluent" in all experiments; Experiments Four and Five suggest that prosodic information in the repair plays an important part; the results of Experiment Three, where listeners were able to detect disfluency before they had recognised the first word of the repair, also allow the possibility that phonetic or acoustic cues other than prosody were of use. But precise definitions of what the prosodic and phonetic or acoustic cues are can only be found by detailed analysis of the speech signal.

In this chapter we look at results of a series of acoustic, phonetic and prosodic analyses of the stimuli used in Experiments One to Five and attempt to establish what cues were used by subjects in making their judgements. We compare our results to those of previous studies of the characteristics of the speech signal in repaired speech and suggest possible areas for future research.

We can distinguish three domains in a disfluent utterance where cues may be

identified: the reparandum, the editing phase and the repair.

From the experiments, there is little evidence that a signal is present within the **Reparandum**. Experiment One explicitly examined this question: the results showed that in the absence of pauses at the interruption point subjects found no indication in the end of the reparandum that the continuation would be discontinuous (such pauses were included in the gate which contained the last word of the reparandum in this experiment) . Where the last word of the reparandum was an incomplete word, subjects were able, in some cases, to identify oncoming disfluency immediately, but it was not always clear that subjects had identified a fragment as such and in principle it would not normally be possible for a listener to know that an incomplete word was present until the onset of the following word at the earliest.

Pauses form part of the **Editing Phase**. They were found to have an effect on disfluency judgements, eliciting more judgements of "disfluent" in all experiments. But the occurrence of a mid-clause pause is not in itself sufficient to positively identify an *overt repair*: mid-clause silent and filled pauses are disfluencies themselves and often occur with no overt repair; false starts and repetitions are not always accompanied by pauses at the interruption point. Another pausal phenomenon, lengthening of the end of the last word of the reparandum, adds an element of fuzziness to the boundary between the reparandum and the editing phase: it is not easy to distinguish where the reparandum ends and the editing phase begins. Discourse markers, or lexical fillers, such as "well" and "I mean" have usually been viewed as being a part of the editing phase, but in our analysis we prefer to view them as being markers of the onset of the repair (see Chapter 3).

Subjects were able to identify most disfluencies in our set of stimuli at an early stage of the onset of the **Repair**. We assume that the cues present at this point are to be found by examination of how the onset of the repair relates to the offset of the reparandum and the editing phase.

We restrict our analysis to a very limited set of data: the spontaneous disfluent stimuli used in the experiments and their spontaneous fluent and rehearsed controls. While this is only a small set of rather heterogeneous data, it is unique in that we have empirical evidence to identify detection points for the disfluencies

to within 35ms. The results of the analyses are not likely to generate answers to the problem of how disfluencies in general are detected, but will show what cues were used by subjects to recognise the disfluencies presented in the experiments and hopefully provide pointers to promising areas for future research.

The analyses will be similar to previous studies, in that we will look for acoustic and prosodic cues in the vicinity of the disfluent interruption, but will differ from other studies in the types of fluent controls that are used for comparison. Nakatani and Hirschberg (1993a,1993b), seeing pause as the major indicator of the presence of repair, compare the acoustic and prosodic information in the vicinity of fluent pauses with that surrounding disfluent pauses. The SRI study (Bear *et al.*, 1992; Shriberg *et al.*, 1992) compare the features of wrongly-hypothesised repairs ("false positives") with those of real repairs. In our study, we compare utterances containing repairs with structurally and prosodically similar fluent utterances. We make use of three sources of data:

1. spontaneous disfluent utterances, used as test stimuli in the experiments;

2. rehearsed fluent versions of the spontaneous disfluent utterances, produced by the method described in Chapter 4, used as control stimuli in Experiments One to Three – in most cases, the speech up to the point where the interruption lay in the disfluent version contained the same words as in that sentence;

3. spontaneous fluent utterances matched with the disfluent set for structure and fluent prosody as far as possible, also used as control stimuli in the experiments.

We begin by looking at the cues examined by other authors (Chapter 2, Section 2.3.2): pauses and F_0 values at the interruption and glottalisation in the reparandum. Then we look at other possible cues: rhythmic factors and word boundary phenomena. The status of the phenomena examined as cues to the presence of disfluency will be discussed with reference to the results of the 35ms gating experiments.

In all the experiments, pauses were taken to manifest discontinuity in the speech signal. In this section we look more closely at the role of pausing in repair

and in the detection of repair.

In analysing the F_0 values we examine a hypothesis based on the premise that pitch declines gradually from the beginning to the end of an utterance under normal circumstances (Pike, 1945). It has been observed, both by Levelt (1984) and in this thesis (Chapter 3), that, where the nature of the disfluency allows, it is possible to excise the reparandum and produce a natural-sounding fluent utterance: the pitch of the repair seems to be reset to an appropriate level for it to link up with the speech which preceded the reparandum, rather than following the route of gradual declination which would be expected with fluent speech. In our analysis of F_0 values we therefore examine the *Reset Hypothesis:* normal pitch declination is stalled in repaired utterances: as a result the F_0 values on either side of a disfluent interruption will not show as great a decline as F_0 values in similar places in fluent utterances.

It has been found for fluent, read speech that the timing of stressed syllables is predictable (Buxton, 1983). Martin (1979) found that disturbed rhythm affected reaction time to phonemes in nonsense strings. If listeners are similarly sensitive to the placing of stressed syllables in spontaneous speech, then a possible cue to the presence of disfluency may be the accompanying disruption to the stress pattern. The subject of isochrony in spontaneous speech is controversial as is the measurement of the points in time at which stress is perceived but in our analysis we make an attempt to judge subjectively whether the stress following the interruption is earlier or later than predicted by the context, rather than attempting precise measurements.

Glottalisation, particularly in vowel-final fragments at the end of reparanda, has been suggested as a possible cue. In the analysis we look for occurrences of glottalisation in our stimuli and discuss their effect on disfluency judgements.

Experiment Three showed that there was often enough information in the first 100ms of the word after the interruption for subjects to detect repair. This may have been too early for F_0 to have been an effective cue where the onset of the repair began with a consonant (although Cardozo and Ritsma (1965) show that listeners are sensitive to changes in pitch within 30ms of their occurrence). Analysis of the speech signal at the onset of the repair will seek acoustic-phonetic and phonological indications that repair has occurred.

7.2 Method

A total of 90 utterances were examined, as detailed above, 30 spontaneous disfluent, 30 spontaneous fluent and 30 rehearsed fluent, which were fluent versions of the spontaneous disfluent set. All had been used as stimuli in the preceding experiments.

All stimuli were sampled at 16kHz and analysed on a Sun Sparcstation, using the Entropic ESPS/Xwaves+ software.

Pauses were defined as periods of apparent silence in the middle of utterances which did not merely coincide with consonantal stop closure or glottal closure before a word beginning with a vowel. They were measured by marking the points on the waveform at the offset of the acoustic signal where the pause began and the point where the onset of the continuation of the signal could be detected by visual and auditory examination. Where restart commenced with a stop consonant, the duration measurement of the pause allowed for a 50ms closure phase.

F_0 measurements were taken on the word prior to and the word following the interruption point in disfluent utterances and at equivalent points in the fluent controls. Two values were taken on each of these words: the peak for the word and the value at the closest measureable point to the interruption (i.e. the last value in the word before the interruption and the first value in the word after the interruption where regular voicing was present).

Other observations were made by close examination of the waveform, pitch-track and spectrogram and by playback of the signal.

7.3 Results

7.3.1 Pauses

Of the 30 disfluent utterances, 14 were found to have no silence at all between the end of the reparandum and the onset of the repair. For the remaining 16 cases, pause-lengths ranged from 34ms to 1134ms. The mean pause-length was 148ms for all disfluent utterances and 278ms for all paused disfluent utterances.

Four pauses were shorter than 100ms. Only two of the eight reparanda ending with fragments had a pause before the onset of the repair. (Shriberg *et al.* (1992) find pause greater than 60ms in 49 of 50 fragments randomly selected from their corpus.) The rehearsed fluent versions of the disfluent utterances contained no pauses at the control points. One pause, of 343ms, was found at a control point in the spontaneous fluent utterances.

The fact that rehearsed fluent versions of the disfluent utterances had no pause at all at the point matched with the interruption point suggests that the structural, phonotactic and prosodic patterns of the disfluent utterances did not predict the occurrence of pause there, as they might in the case of fluent pauses between clauses or sentences (Gee & Grosjean, 1983).

To test for the structural and prodosic expectancy of pauses at the interruption points more formally, Gee and Grosjean's Φ algorithm (1983, based around Selkirk, 1980) was applied to the 16 disfluent utterances that contained silent pauses. In two utterances the interruption was mid-word, where no pause is expected. In all but one of the remaining utterances, the pause occurred in the middle of a phonological phrase (ϕ-phrase), rather than at a clause boundary, where pause would not be expected in normal, fluent speech. In the remaining case, the pause was at a phrase boundary, where a pause was more likely to occur in fluent speech (although the rehearsed fluent version of this utterance had no pause).

Now we turn to the question of how subjects reacted to pauses in the experiments. It is most interesting to look at the responses from Experiment Three (35ms gating experiment), since, with the stimulus gradually incrementing in length on each presentation, we can assess where subjects began to feel that the pause was becoming too long to be fluent or whether other factors such as breath played a part in the results.

For the 16 disfluent utterances with a pause at the interruption, close visual and auditory examination of the waveform revealed that 11 of the pauses contained no noise above the background level, 2 contained an audible inhalation and 3 were accompanied by other very brief audible vocal sounds.

In the cases where there was no sound at all in the pause, 5 contained pauses too short for any gradual increase in mean disfluency judgements to be easily

observed: these were pauses of 105ms or less, which would only stretch over a maximum of 3 whole 35ms gates. Of the remaining 6, 4 show a gradual rise over the gates containing the pause (see figures 7.1 and 7.2 for examples), one has a fragment ending in falsetto phonation, the possible effect of which cannot be separated from the possible effect of the pause (the mean disfluency judgements rise steadily through the silent section) and the remaining one also has a fragment-final reparandum, but has no rise in mean disfluency judgements until the onset of the repair (figure 7.3). With such a small amount of relevant data, it is difficult to make claims about how long the pause had to be before subjects began to react to it by changing their judgements from "fluent" to "disfluent", and even with more data, the value of such claims would be dubious, since it may be that subjects just became less certain about the disfluency judgement task as they heard three or more apparently identical presentations as the pause developed. We can observe from figures 7.1 and 7.2 that the gradual increase in mean disfluency judgements begins fairly early in the pause in these two cases, but increases sharply when the onset of the following word is heard.

Where the pause contained an audible inhalation, mean disfluency judgements rose steeply at the point where the breath was perceived (figure 7.4). Whether such a cue is perceptible in normal listening situations as opposed to the listening booth, where the speaker is in effect only a short distance from the subjects' ears is a moot point. The same applies to the other sounds which were found on close examination of the signal. One example of these is illustrated in figure 7.5: at the offset of voicing in the bilabial closure at the end of the word *"some"*, there is a glottal stop with bilabial closure. This sound can not really be described as marking an *abrupt* cut-off in the signal, since the word it terminates is quite fully pronounced. It may be a feature of 35ms gating presentation of high-quality digitised recordings of speech that such minor phonetic events become exaggerated, where in normal listening conditions they would pass unnoticed.

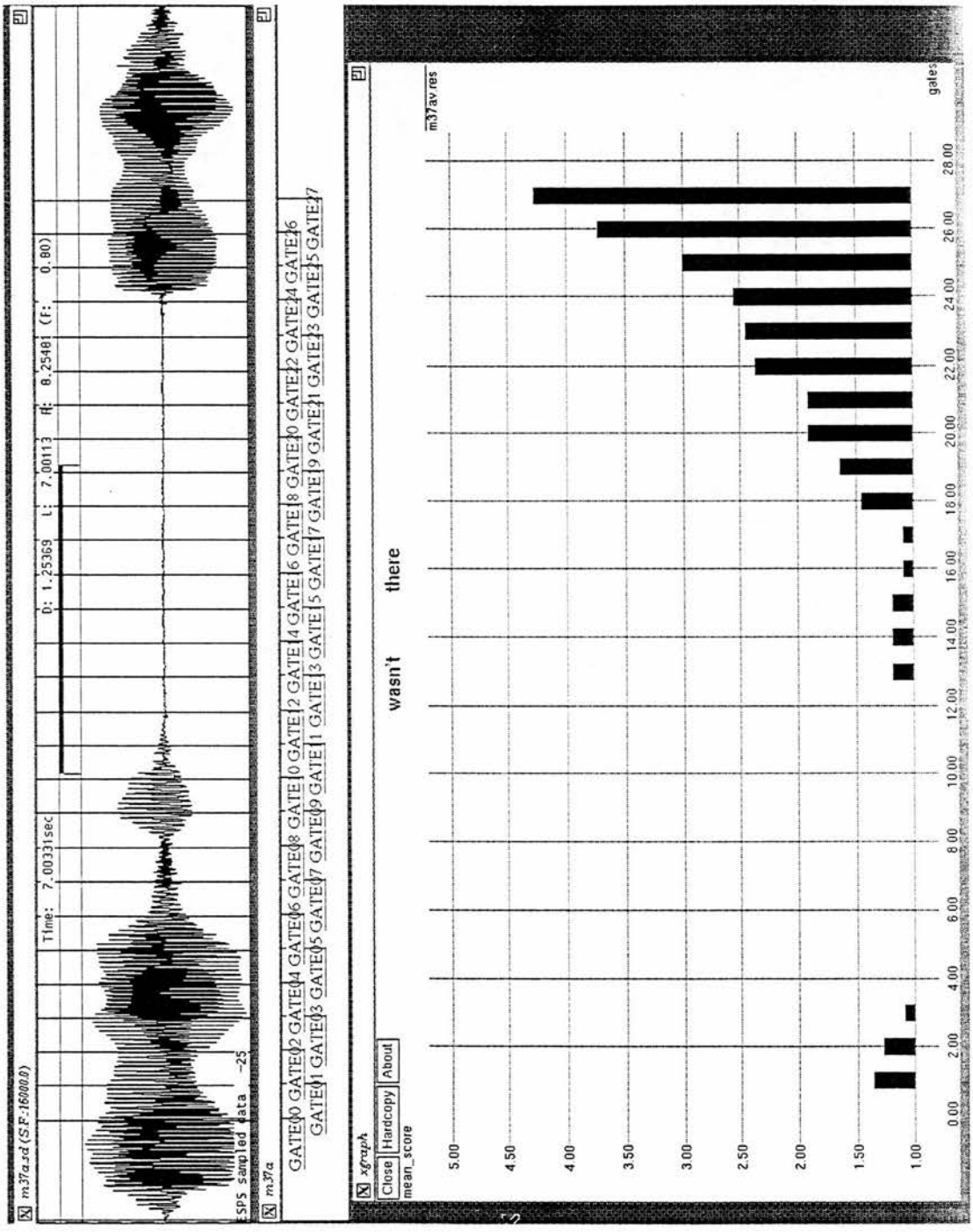

Figure 7.1. Acoustic Analysis: section of waveform and mean fluency judgements at each gate in Experiment 3 for "there **wasn't** — **there** wasn't a great deal of choice". Pause length at interruption = 405ms. Fluency Judgements: 1="fluent", 5="disfluent". More "disfluent" judgements as pause lengthens.

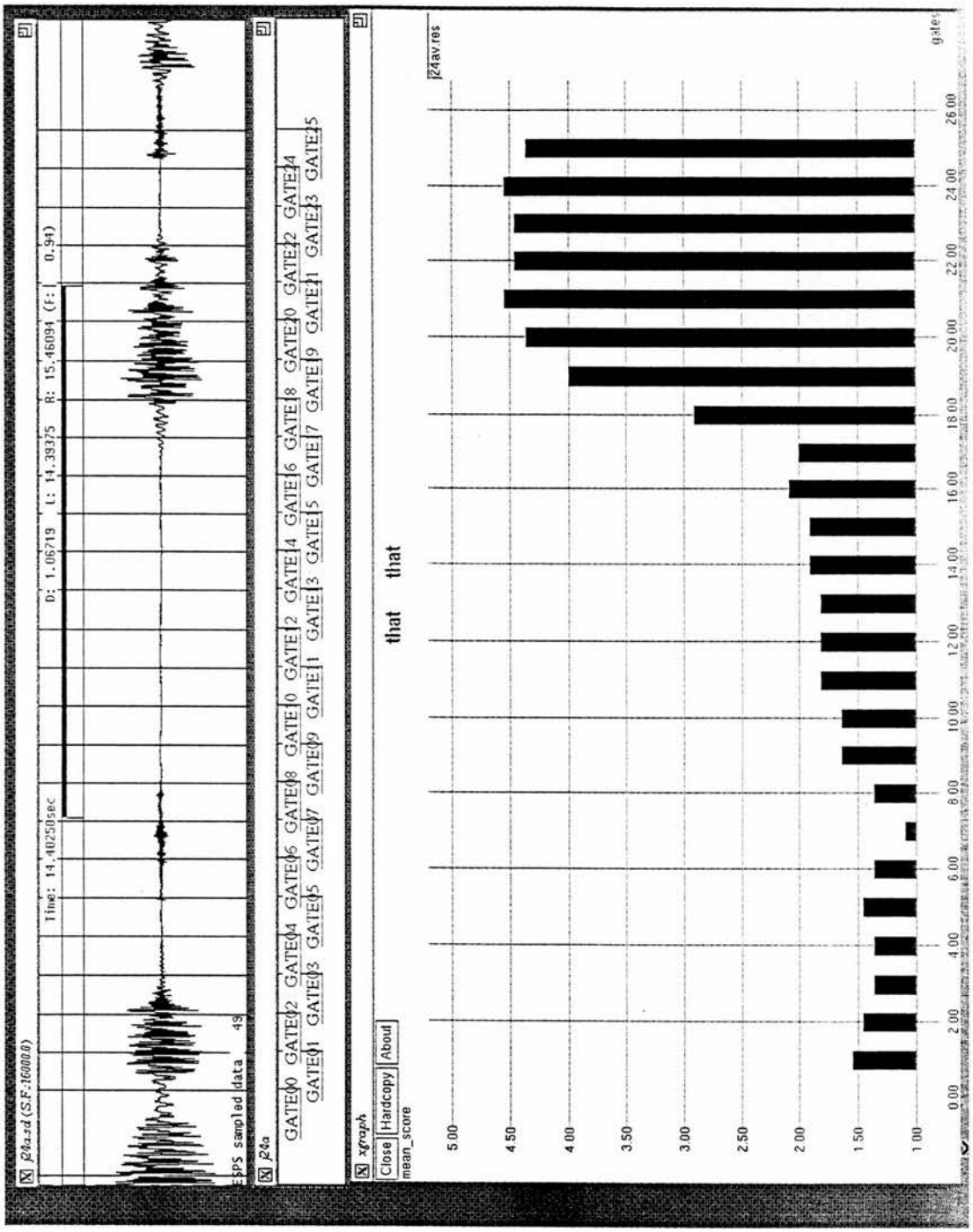

Figure 7.2. Acoustic Analysis: section of waveform and mean fluency judgements at each gate in Experiment 3 for "they've thrown away **that** — **that** trump card". Pause length at interruption = 288ms. Fluency Judgements: 1="fluent", 5="disfluent". More "disfluent" judgements as pause lengthens.

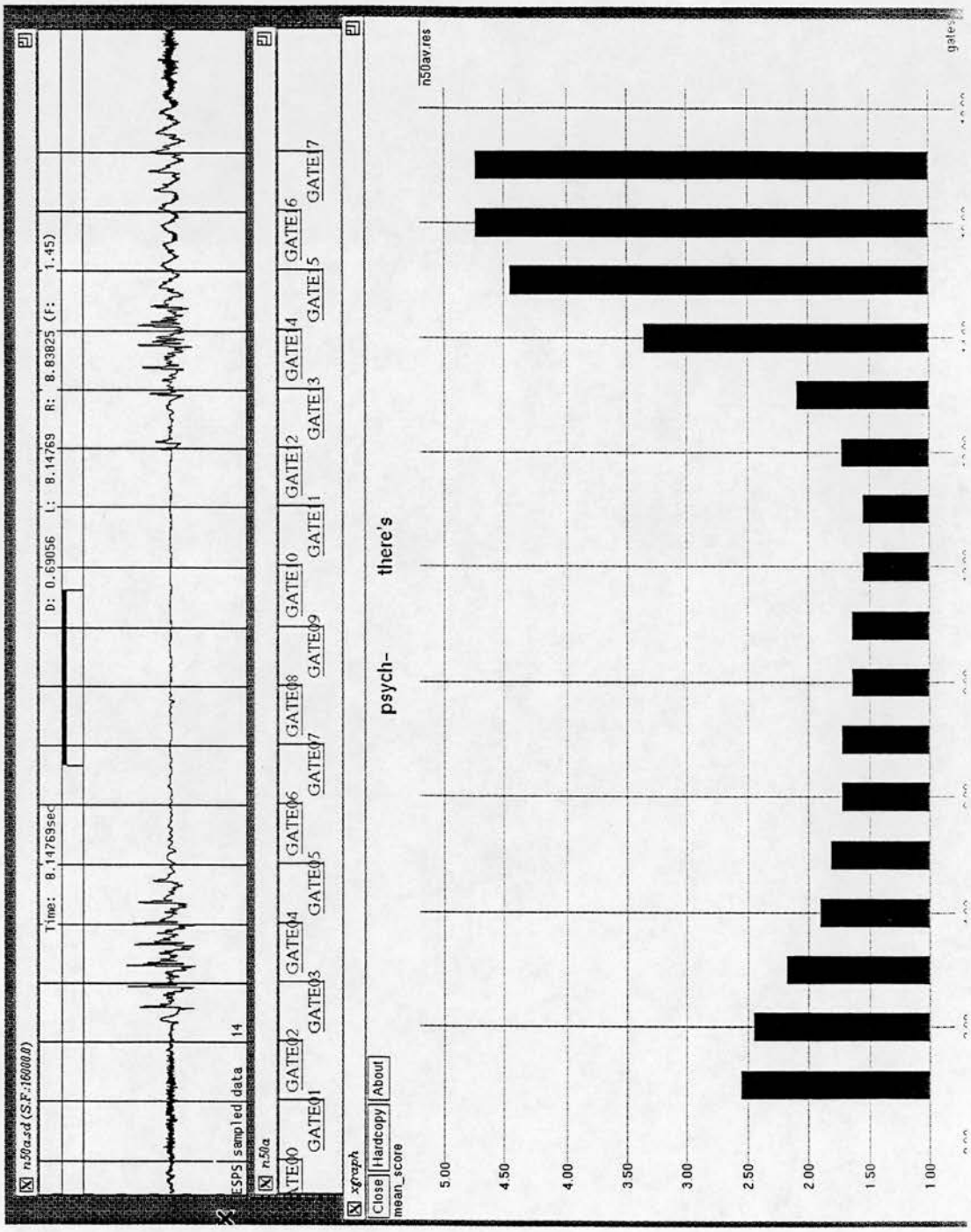

Figure 7.3. Acoustic Analysis: section of waveform and mean fluency judgements at each gate in Experiment 3 for "one of the things I thought the **psych— there's** a psychologists' meeting ...". Pause length at interruption = 210ms. Fluency Judgements: 1="fluent", 5="disfluent". No rise in "disfluent" judgements until repair onset.

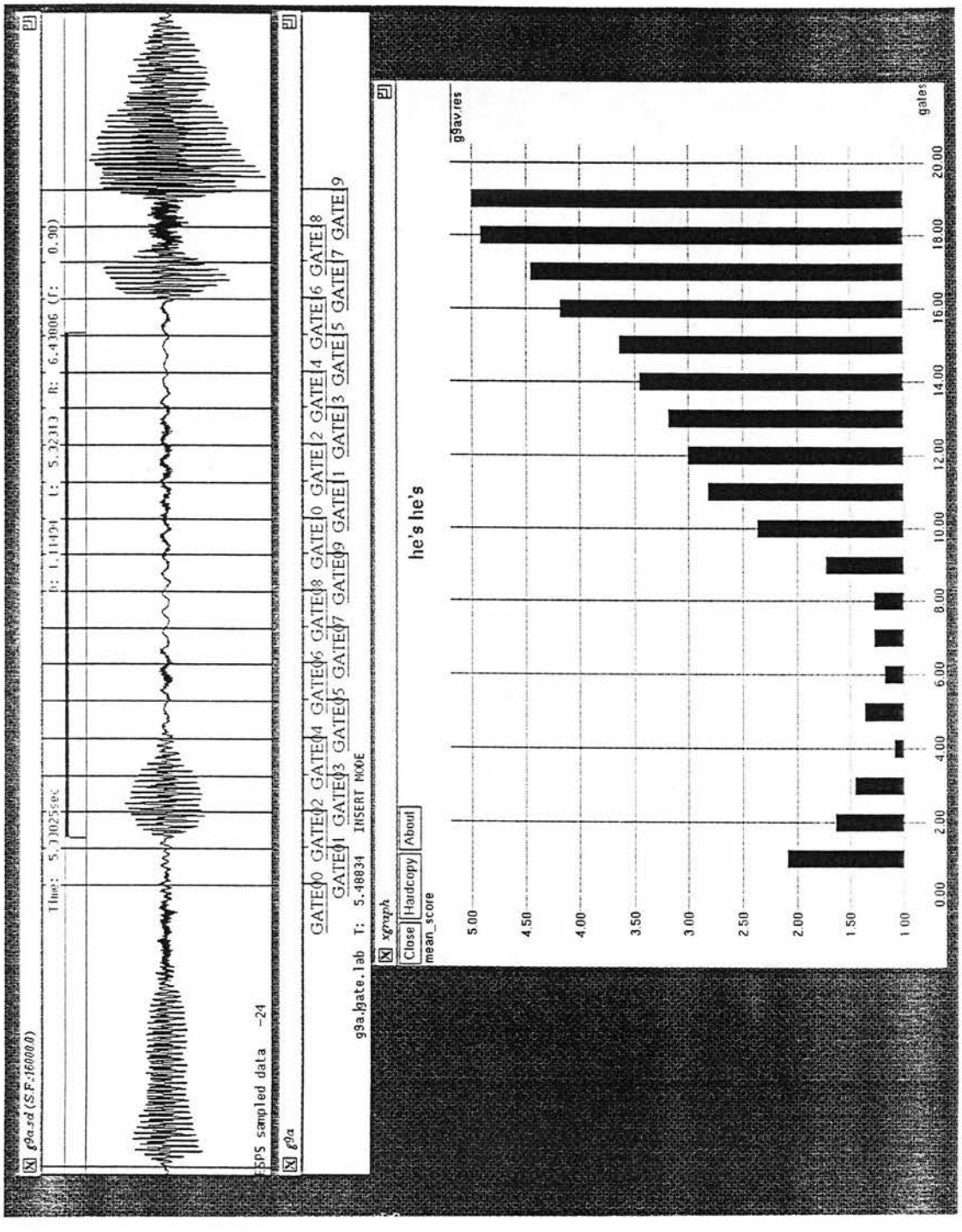

Figure 7.4. Acoustic Analysis: section of waveform and mean fluency judgements at each gate in Experiment 3 for "it's quite obvious **he's** — **he's** on something". Pause length = 287ms. Fluency Judgements: 1="fluent", 5="disfluent". Steep rise in "disfluent" judgements at onset of inhalation, gates 9-10.

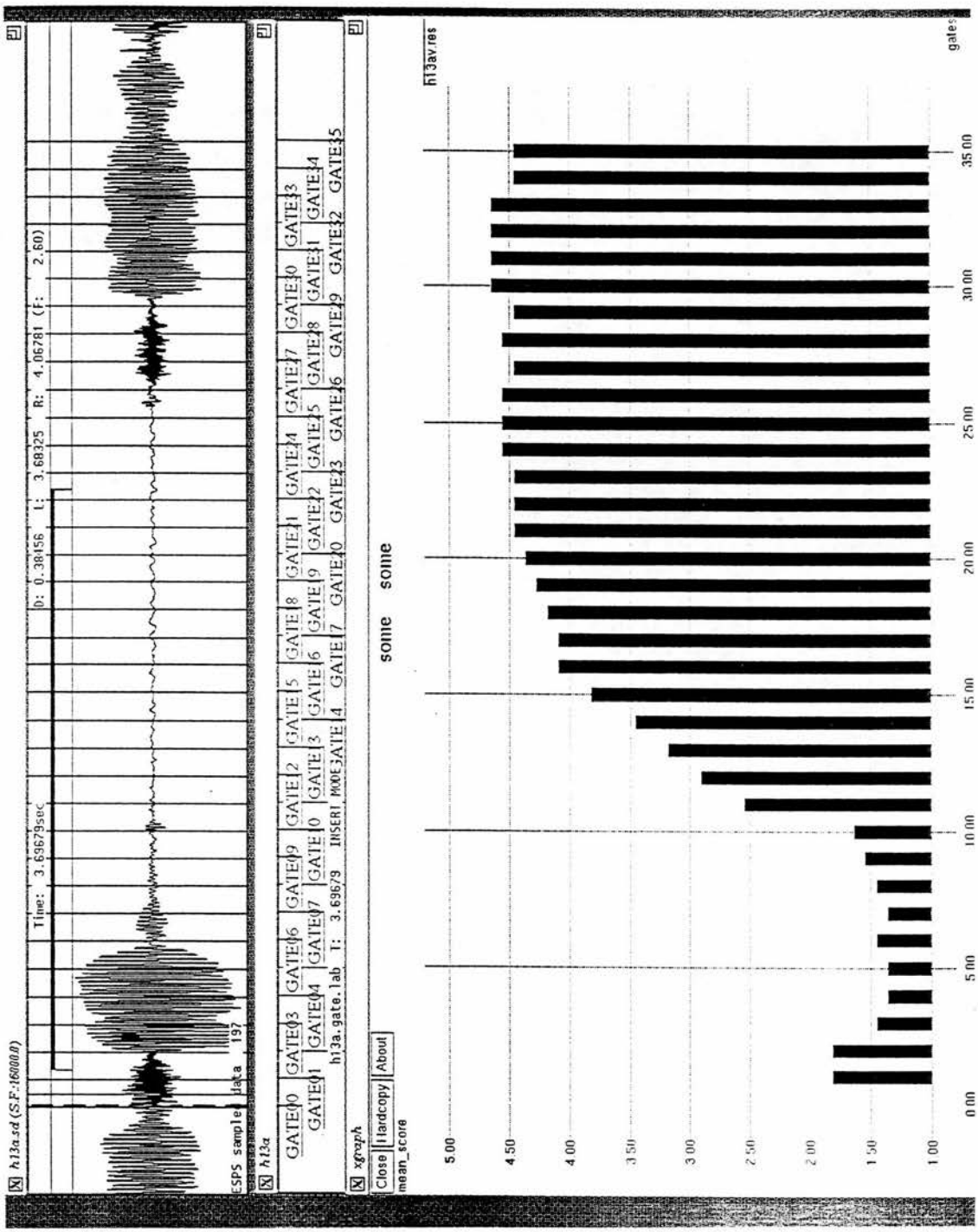

Figure 7.5. Acoustic Analysis: section of waveform and mean fluency judgements at each gate in Experiment 3 for "but in **some** — **some** English universities ...". Fluency Judgements: 1="fluent", 5="disfluent". At gate 11, at the offset of the voiced bilabial closure phase, a glottal stop with bilabial closure is heard, which prompts "disfluent" responses.

7.3.2 Pitch

In analysing the intonational characteristics of repairs we compared the F_0 measured in Hz at selected points before and after the interruption in disfluent utterances with similar points in the matched controls. These comparisons were used to test the **Reset Hypothesis**, which predicted that F_0 would be reset after the interruption to a level higher than would be expected if the utterance had continued fluently.

For the set of all repairs, no significant differences were found between the disfluent and fluent utterances for F_0 differences between pre-interruption offset and post-interruption onset values. A significant difference *was* found for the difference in peak F_0 values for the comparison of the disfluent utterances with the spontaneous fluent controls: there was found to be a greater fall in F_0 over the two points examined in the fluent cases than in the disfluent cases ($t = 3.69$, $df = 29$, $p < 0.001$). The word after the interruption in the disfluent utterances was an average of 5.9Hz lower than the word before, while the average difference for the fluent utterances was -56.7Hz. But in the rehearsed fluent versions of the disfluent set, the fall in F_0 across the two points (18.7Hz) was not significantly greater than that found in the disfluent set. These results only offer partial support for the reset hypothesis for the set of all the stimuli.

It was possible that different results could emerge for different types of disfluency. For this reason the tests were repeated for the separate sets of false starts, repeats and fragments (irrespective of whether they were false starts or repeats). The only significant difference that was found for these comparisons was for false starts: in 9 of 11 cases, peak values of F_0 fell more in spontaneous fluent utterances ($\bar{X} = -46.96Hz$) than in their disfluent pairs ($\bar{X} = -5.43Hz$) ($W = 10, N = 11, p < 0.05$).

Next, the disfluent stimuli were examined further with analyses of their syntax and prosodic structure.

Structural analysis of the false starts showed that in 9 of the 11 cases the restart was sentence-initial. These restarts had typical sentence-initial prosody (high F_0 and intensity) but were not always higher in F_0 than the preceding peak (in the reparandum), because in most cases the preceding peak was itself

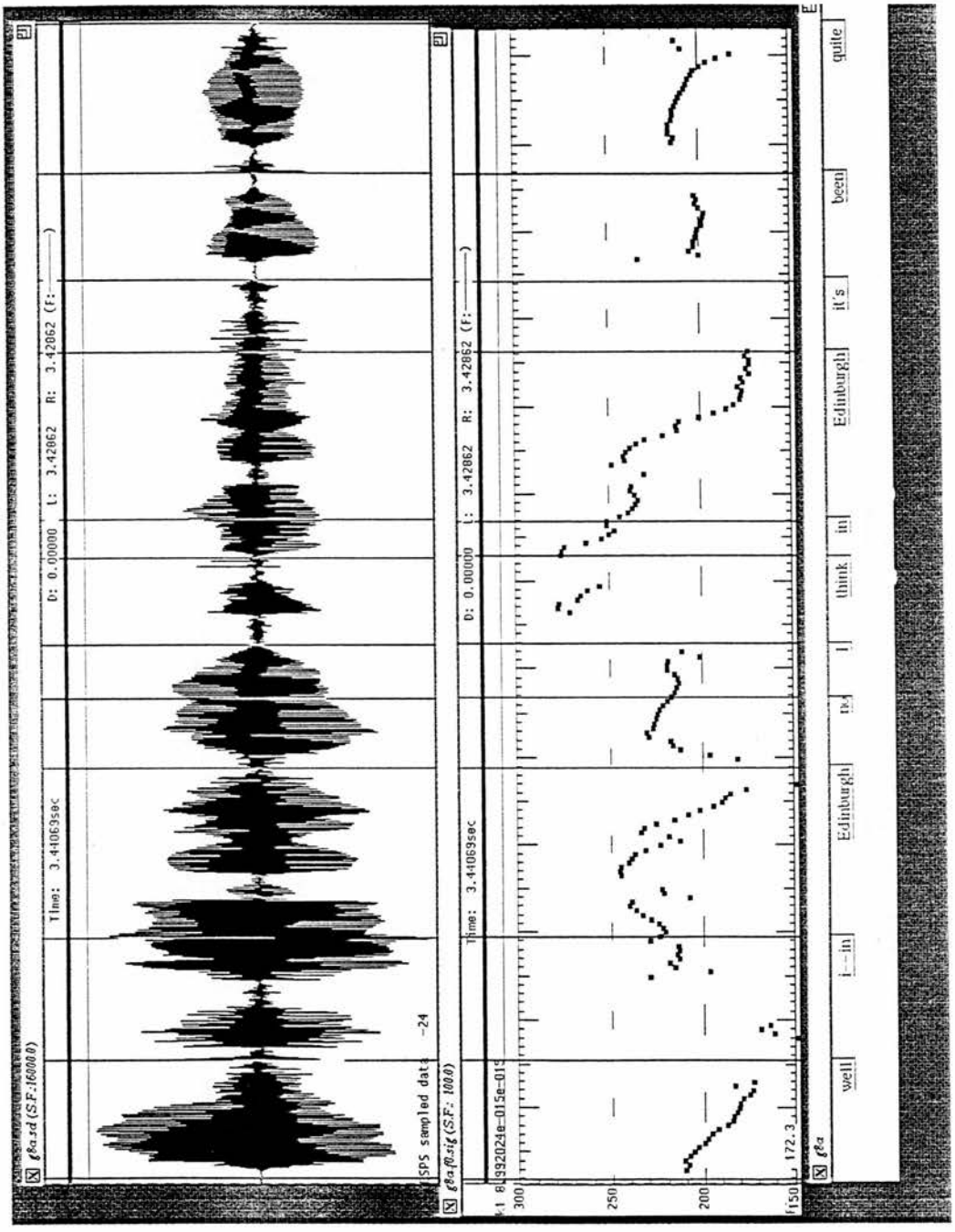

Figure 7.6. Acoustic Analysis: section of waveform and pitch-track for utterance with high pitch on repair: "no I THINK in EDinburgh ...".

sentence-initial: the sentence-initial stressed syllable had higher F_0 in the restart than in the original utterance in 4 cases, and was lower in 3 cases (see example of high restart in figure 7.6). But the importance of the restart's sentence-initial prosody from the perceptual point of view is not in how it compares to the prosody of the reparandum, but in the fact that it is *unexpected.* The intensity and F_0 associated with sentence onset and initial peak is rarely found in unmarked speech within the sentence: if it is perceptually distinct from other recipients of high intensity and F_0, such as heavy emphasis and contrastive stress, or if such other cases are predictable on the basis of context to some extent, its occurrence may provide a clear prosodic cue to the listener that the speaker has backtracked and restarted the current sentence. Of the remaining two false starts, one had the repair commencing at the onset of a subordinate clause: in this case it is clear from listening to an edited version of the signal with the reparandum removed that the F_0 of the repair is at a suitable level to link with the speech prior to the reparandum, but it is not possible to dismiss the possibility that a word with the same F_0 could also follow fluently from the end of the reparandum. The repair in the remaining false start is simply a correction to the pronunciation of the word that formed the reparandum (*"valiency valency"*): the F_0 of the repair is close to that of the reparandum, dropping from 211-186Hz as opposed to 216-191Hz in the reparandum.

For repetitions, F_0 was compared for the first and second instance of the repeated word or string. If it were the case that simple retracing occurred in repetition, we would expect to find no change in F_0 values between the first and second instances. The set of 11 repeats included 8 single word repetitions, 2 two-word repetitions and one five-word repetition. Of the single word repetitions, only one involved repetition of a content word, but two others were repetitions of stress-bearing function words. The 2 two-word repetitions both contained just one stressed syllable and the five-word repetition contained two stressed syllables. The F_0 analysis showed that only 3 repetitions demonstrated the sort of pattern expected for simple retracing, with the same pitch on both instances: all three contained stressed syllables (figures 7.7, 7.8 and 7.9). Of the other 3 repetitions containing stressed syllables, two had lower F_0 on the second instance of the repeated section and one had higher F_0 on the second instance. The 5 repetitions

containing unstressed words only (all were single-word repetitions of monosyllabic function words) also varied in their F_0 patterns: 3 contained F_0 rises from the first to the second instance, 1 had a fall and 1 had no voicing. The F_0 of the second instance of the repeated word seemed to be related to that of the following stressed word, which formed the head of its phonological phrase: where the F_0 rose with respect to the first instance, the F_0 of the head was higher, and where it fell, the F_0 of the head was lower.

Analyses of reparanda ending in fragments also showed a variety of outcomes. Of the four fragment-final false starts, one had sentence initial prosody in the restart, one, a word-substitution, had no difference in F_0 between the reparandum and the repair, one showed a slight fall in F_0 and one, also a lexical substitution, showed a sharp rise, exhibiting contrastive stress in the repair (figure 7.10). Of the four repetitions with fragment-final reparanda, two had no change in F_0 between the two instances, one had higher F_0 on the second instance and one had no voicing.

7.3.3 Rhythm

The metrical structure of all 30 disfluent stimuli was examined informally and an estimate was made of the approximate point in time where the next stressed syllable after the interruption would occur if the utterance continued fluently. In cases where the repair was a retracing (with the same structure but one or more altered lexical items) or a repetition of the reparandum, the estimate was made by looking at the metrical structure of the appropriate place in the repair; in other cases it was done on the basis of the best hypothesis for a fluent continuation.

The difficulty and unreliability of the task was reflected in the fact that for over a quarter (8) of the stimuli no decision could be made. Of the estimates that could be made, stress was estimated to be delayed in 11 cases and early in 2; in the remaining 9 cases it was estimated that the time of the stressed syllable after the interruption was within the region where a fluent continuation would also have had stress.

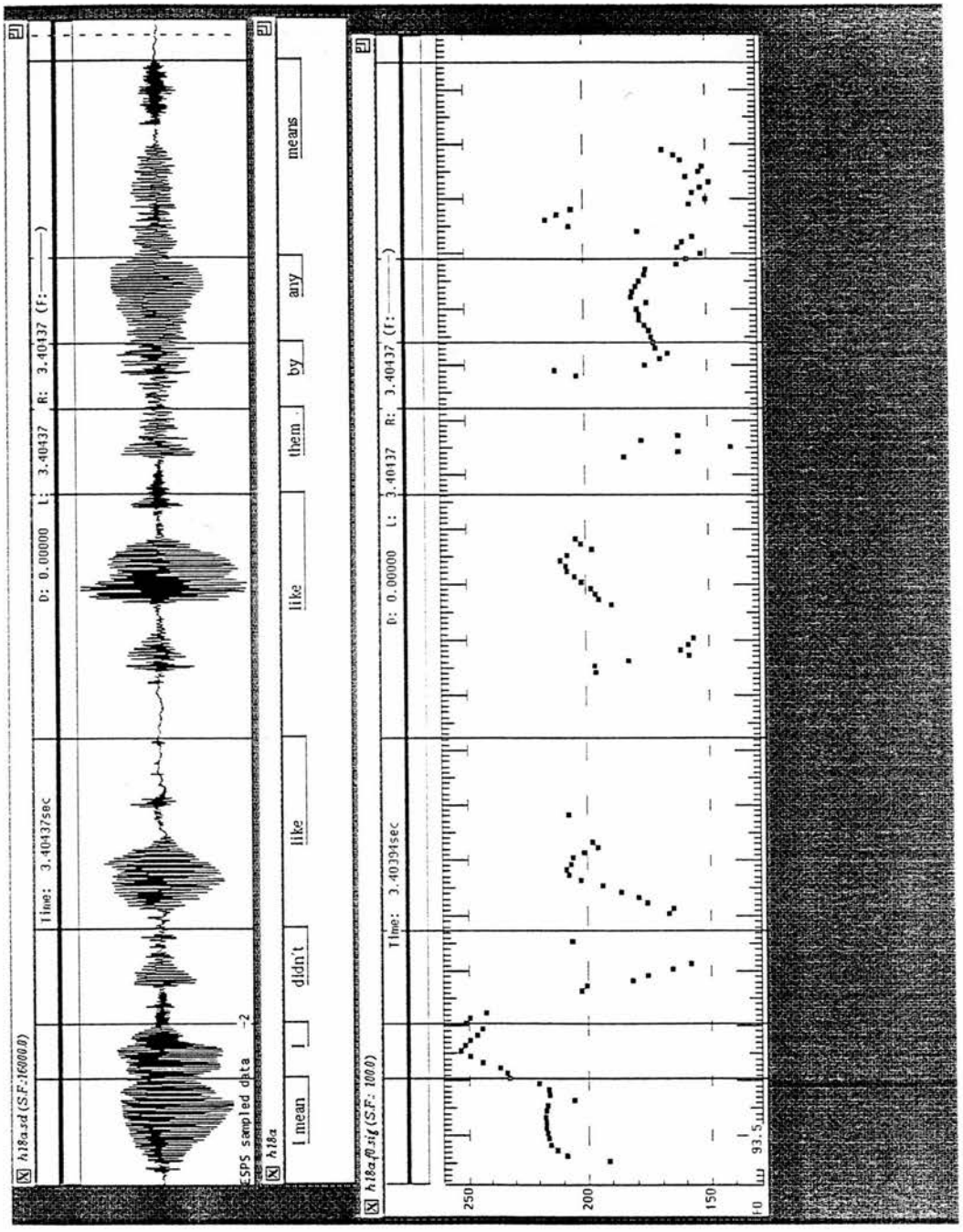

Figure 7.7. Acoustic Analysis: section of waveform and pitch-track for repetition with "repeated pitch".

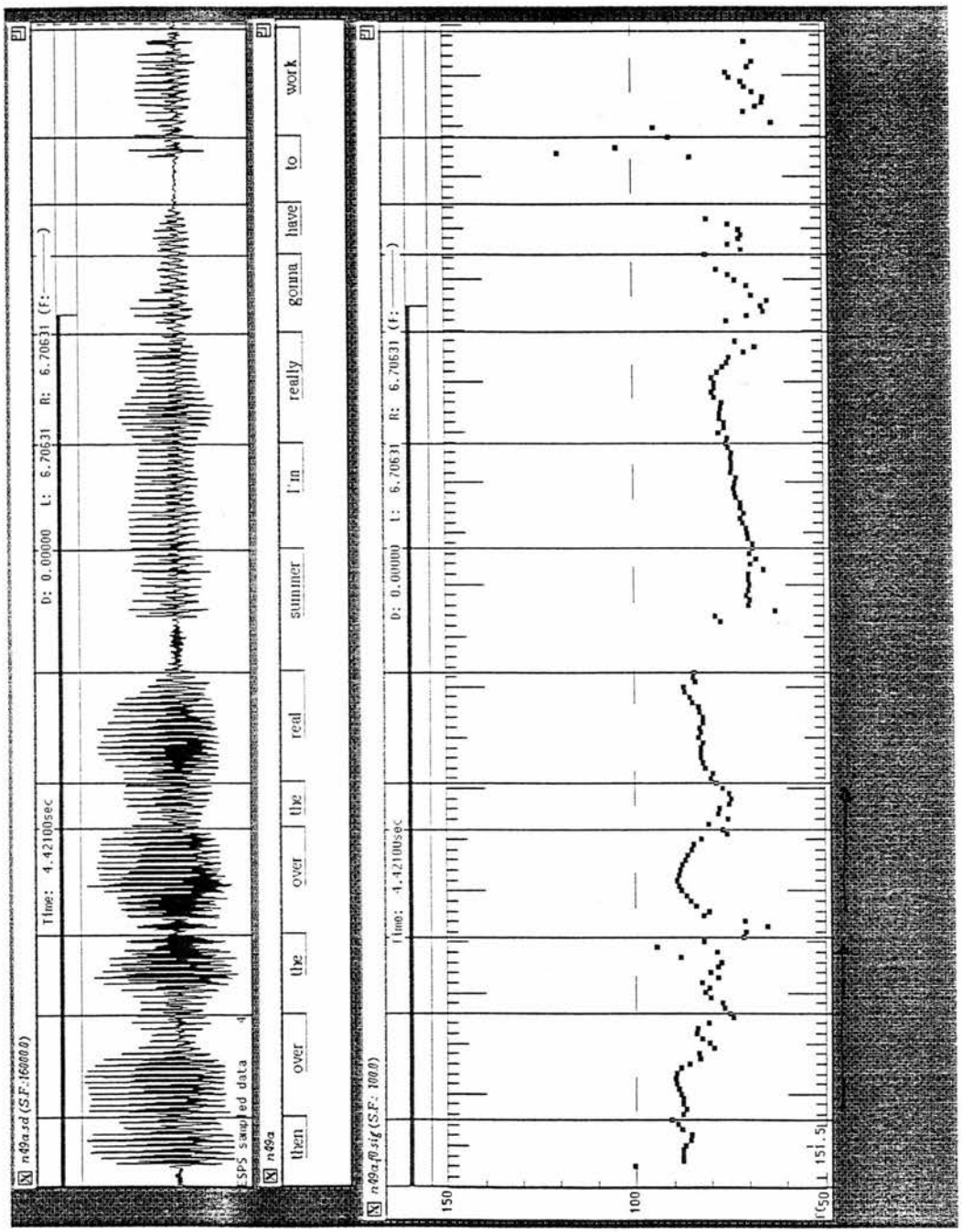

Figure 7.8. Acoustic Analysis: section of waveform and pitch-track for repetition with "repeated pitch".

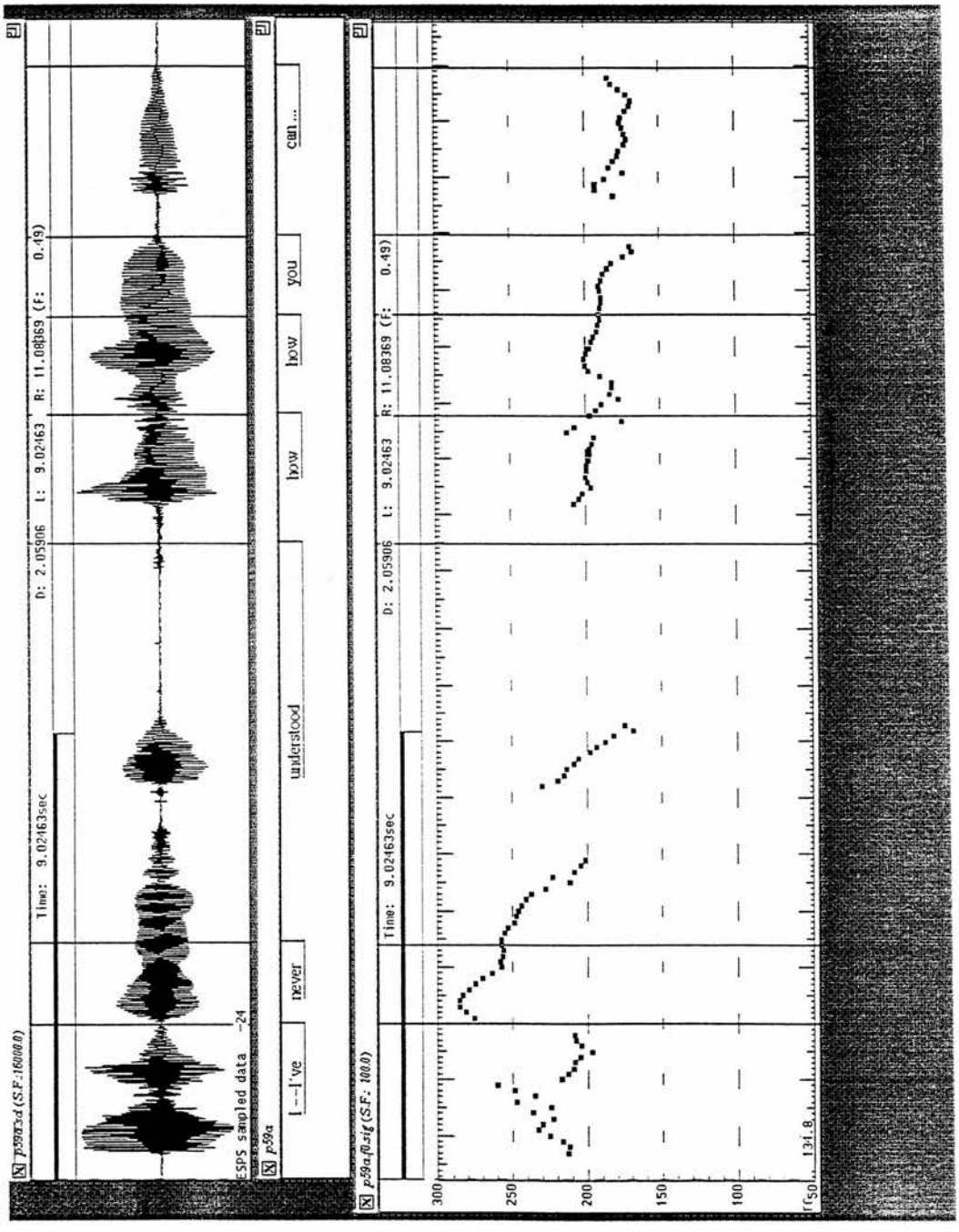

Figure 7.9. Acoustic Analysis: section of waveform and pitch-track for repetition with "repeated pitch".

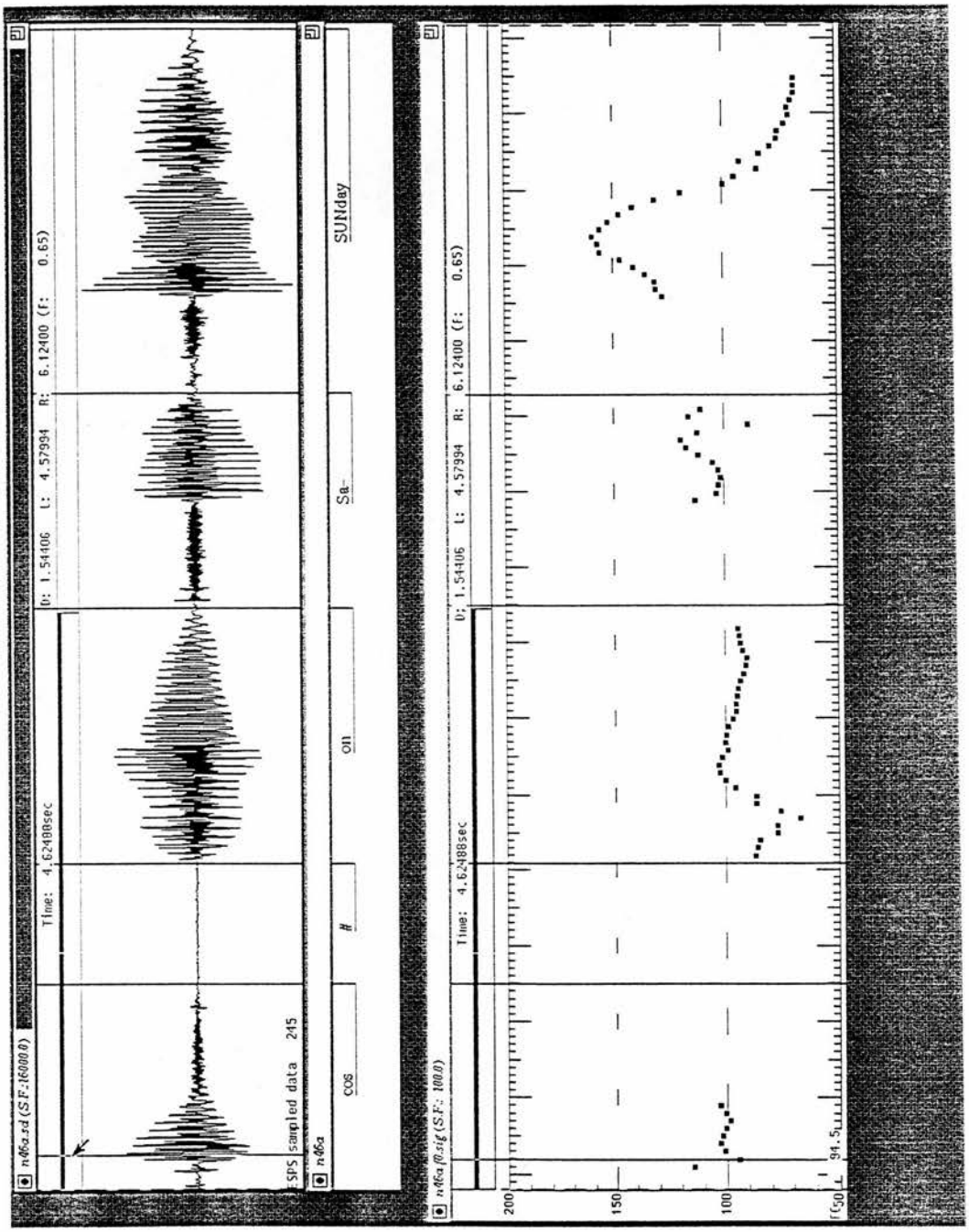

Figure 7.10. Acoustic Analysis: section of waveform and pitch-track for repair with contrastive stress.

7.3.4 Glottalisation

Reparanda were categorised as "glottalised" where they ended in a vocal section with a deceleration in the rate of vibration of the vocal cords or a glottal stop. This deceleration could be caused by adduction of the vocal cords or just by reduction in subglottal pressure: we make no distinction in this study.

In our sample, 9 of the 30 disfluent utterances had a glottalised reparandum offset. For reparanda ending in fragments, 2 of the 5 vocal-final fragments were glottalised; one of the others ended in a low-intensity *falsetto* segment, as if the vocal cords had been tensed rapidly at the onset of the vowel – this is different from the glottalised offsets which were generally slightly less tense and of higher intensity with more irregular pitch pulses. For reparanda ending in full words, 2 of the 11 repeats and 5 of the eleven false starts had glottalisation at the offset. Figure 7.11 provides a clear example. In most cases glottalisation of the offset of the reparandum was followed by immediate repair, with no silent pause: the once exception to this was a very brief (80ms) silence.

In 4 utterances, significant glottalisation was found at the *onset of the repair* (e.g. figures 7.12 and 7.13). In two of these cases, the glottalised onset was preceded by a silent pause of over 150ms and in a third, by lengthening of the last segment of the reparandum.

There are two differences between this small sample and other studies. Firstly, Shriberg *et al.* and Nakatani and Hirschberg focus on glottalisation in fragments: in our data, glottalisation is as common in reparanda ending in full words as it is in fragment offsets. The important factor seems to be whether or not the reparandum ends in a voiced segment. Secondly, Shriberg *et al.* find that pause and glottalisation combine to signal the presence of repair: in our data, glottalisation in the reparandum is not usually accompanied by silent pause.

It is not always a straightforward task to distinguish between interruption glottalisation and the other types of glottalisation (or laryngealisation) that occur in fluent speech. Nakatani and Hirschberg claim that interruption glottalisation is acoustically different from creaky voice at the end of prosodic phrases, glottal stops and epenthetic (intervocalic) glottalisation, but they do not specify how. Shriberg *et al.* found that interruption glottalisation was usually of a

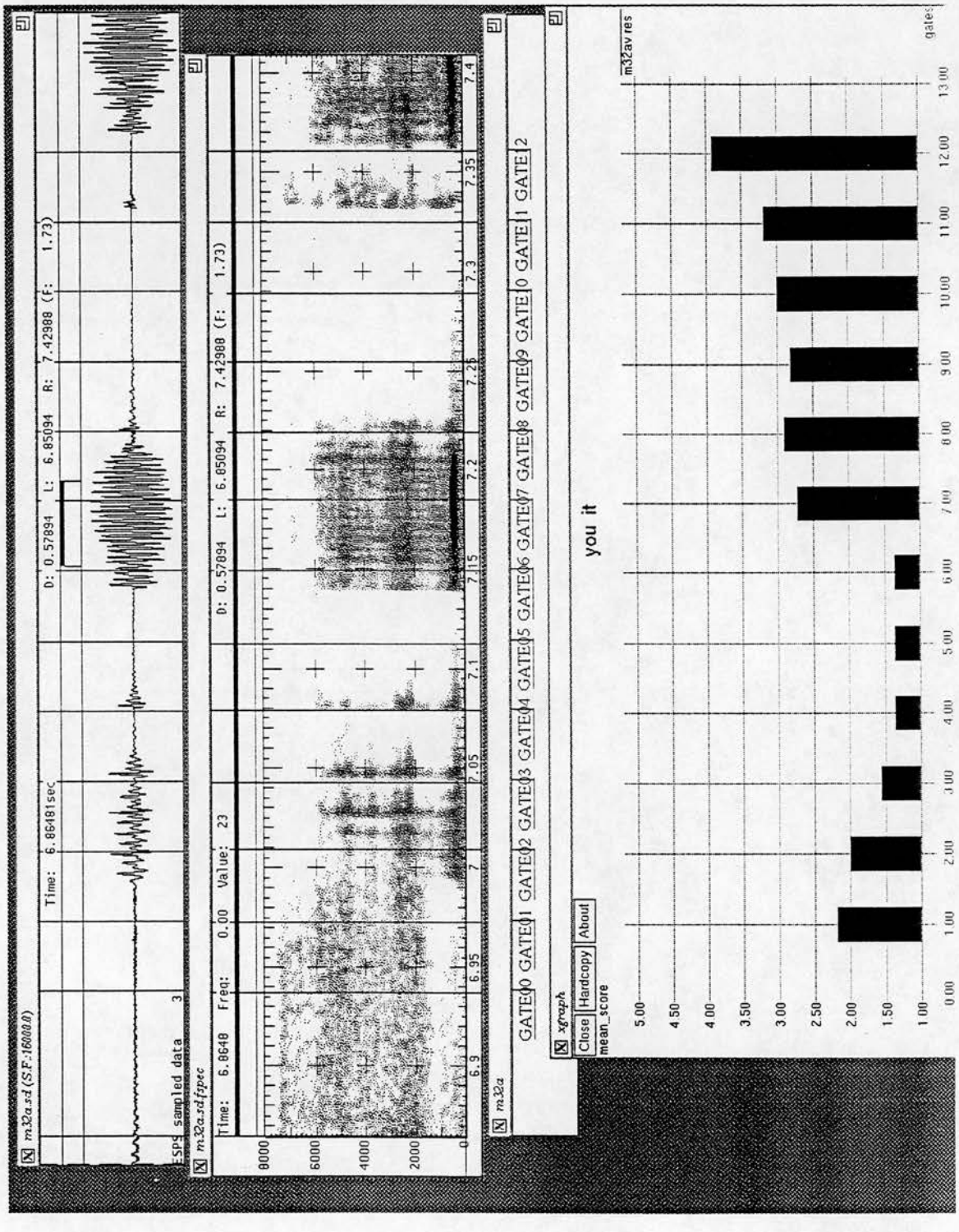

Figure 7.11. Acoustic Analysis: section of waveform and fluency judgements from Experiment 3 for "and if **you** — **it** just ...". No rise in judgements of "disfluent" until onset of repair.

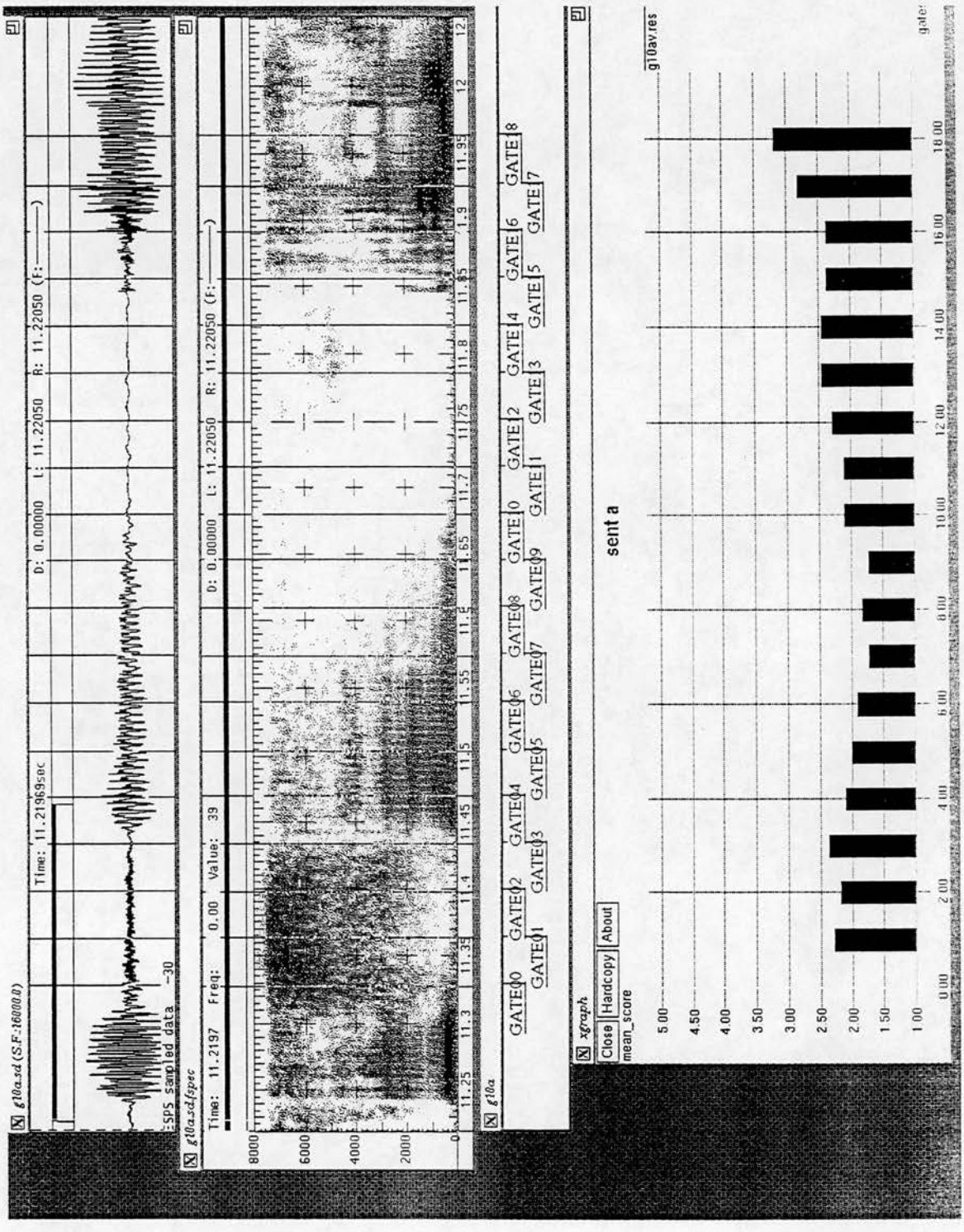

Figure 7.12. Acoustic Analysis: section of waveform and fluency judgements from Experiment 3 for "they **sent** — **a** lot of their youngsters would ...". Glottalised onset to repair.

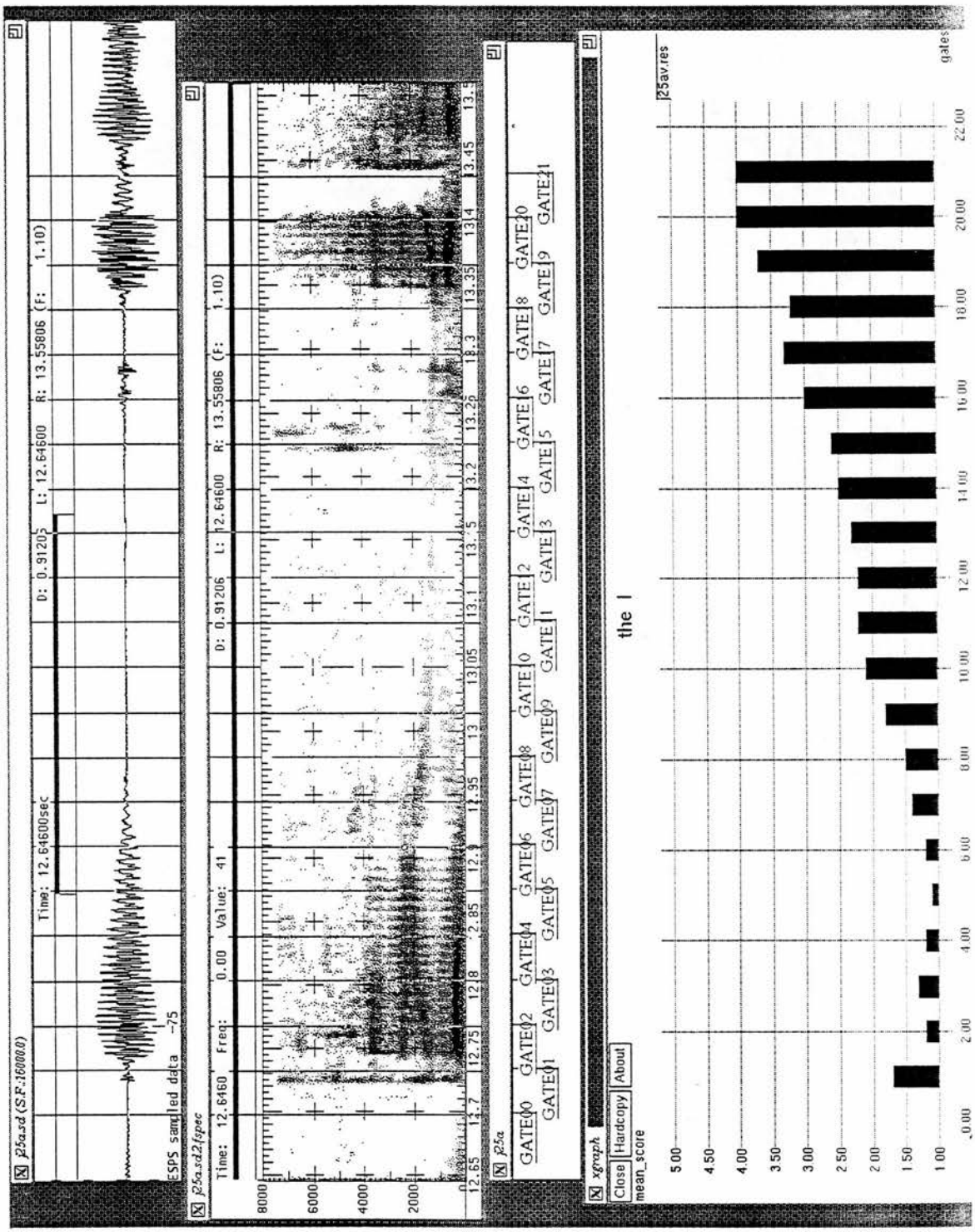

Figure 7.13. Acoustic Analysis: section of waveform and fluency judgements from Experiment 3 for "I don't know what **the** — **I** don't know what the ...". Glottalised onset to repair.

higher intensity than other types. In some cases in our data we can find distinguishing features: the glottalisation at the end of "you" in a disfluent utterance (figure 7.14) is clearly different from that in the word "you" in the same phrasal position in a fluent utterance by the same speaker (figure 7.15) – it shows both greater tenseness of the vocal cords and a slow-down to stop of glottal vibration where the fluent version shows slow but regular vibrations of the vocal cords and laxness characteristic of creaky voice. In other cases, however, it was difficult to distinguish between interruption glottalisation and fluent glottalisation which coincided with interruption: in the case of *"in Edinburgh – no I think in Edinburgh it's been quite ..."* there is word-final glottalisation at the end of both instances of Edinburgh, which may be put down to phrase-final glottalisation; in *"I don't know if it – how true it is"* the first instance of "it" is glottalised throughout, which may just be a feature of laxness, rather than an effect of interruption. It remains to be seen whether there is a clearly definable "interruption glottalisation". It may be simply that greater tenseness in glottalisation is a function of the phrasal position of the interruption: where an interruption occurs in a portion of speech which is characterised by laxness, the associated glottalisation may not be acoustically distinguishable from normal glottalisation.

Comparison of episodes of glottalisation with mean disfluency judgements at the gates at the same points in time did not show any direct link between the two: subjects did not respond immediately to glottalisation by giving more judgements of disfluent, but usually waited until the beginning of the repair. This is not to say that glottalisation did not serve as a cue: it may have contributed with delayed effect, in combination with later cues.

In conclusion, glottalisation is often found where the reparandum ends in a vowel or voiced sound, whether the reparandum ends in a fragment or not. In our data it does not coincide with silent pauses at the interruption point, but with immediate repairs. Some vocal repair onsets also showed glottalisation. It is unclear from the experimental results whether glottalisation had an effect on disfluency judgements, although no immediate effect was observed. It is also unclear whether it is possible in principle to distinguish interruption glottalisation from other forms of glottalisation which occur in fluent speech.

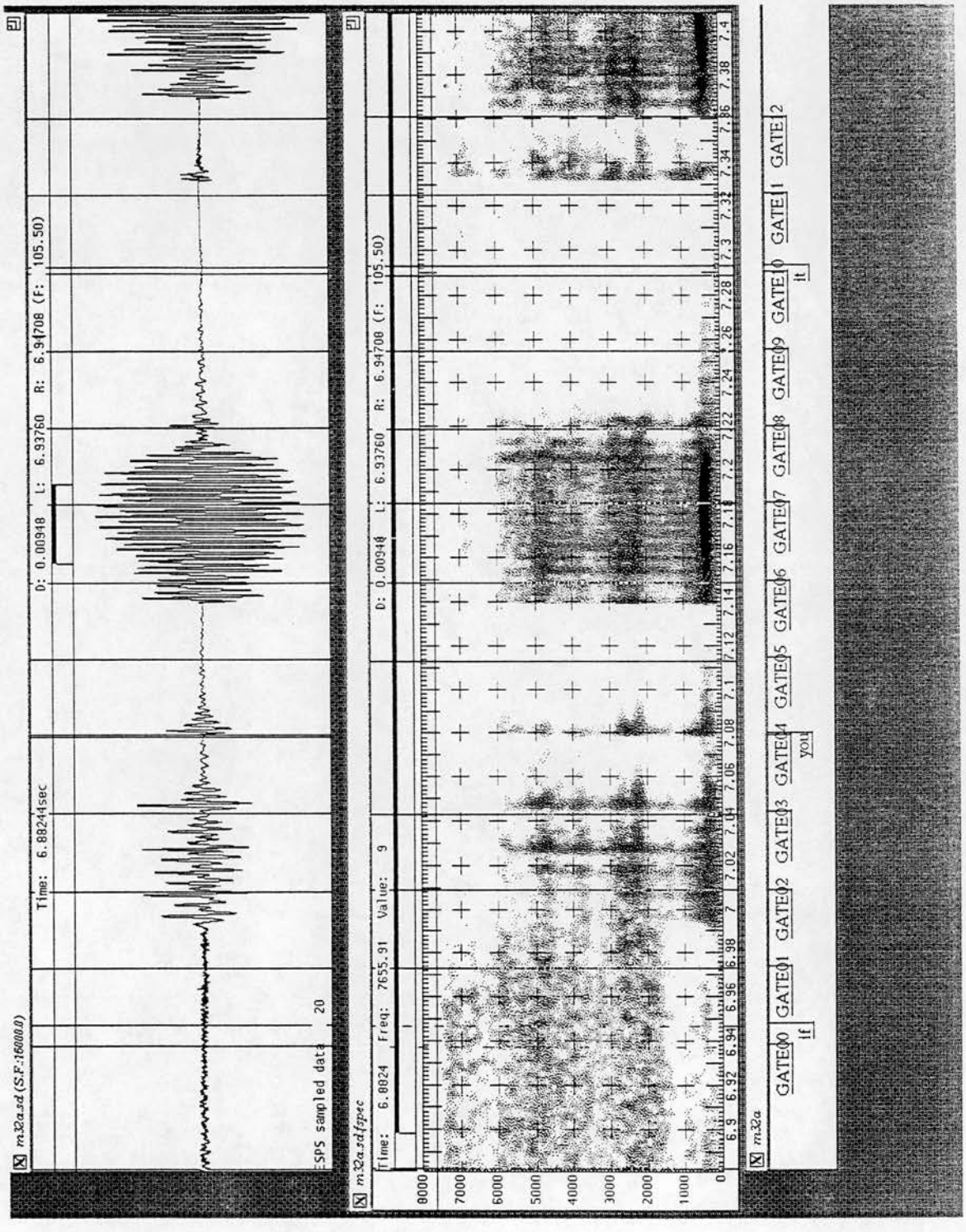

Figure 7.14. Acoustic Analysis: section of waveform and spectrogram and fluency judgements from Experiment 3 for "if **you** — **it** just ...". Glottalisation at interruption to compare with figure 7.15.

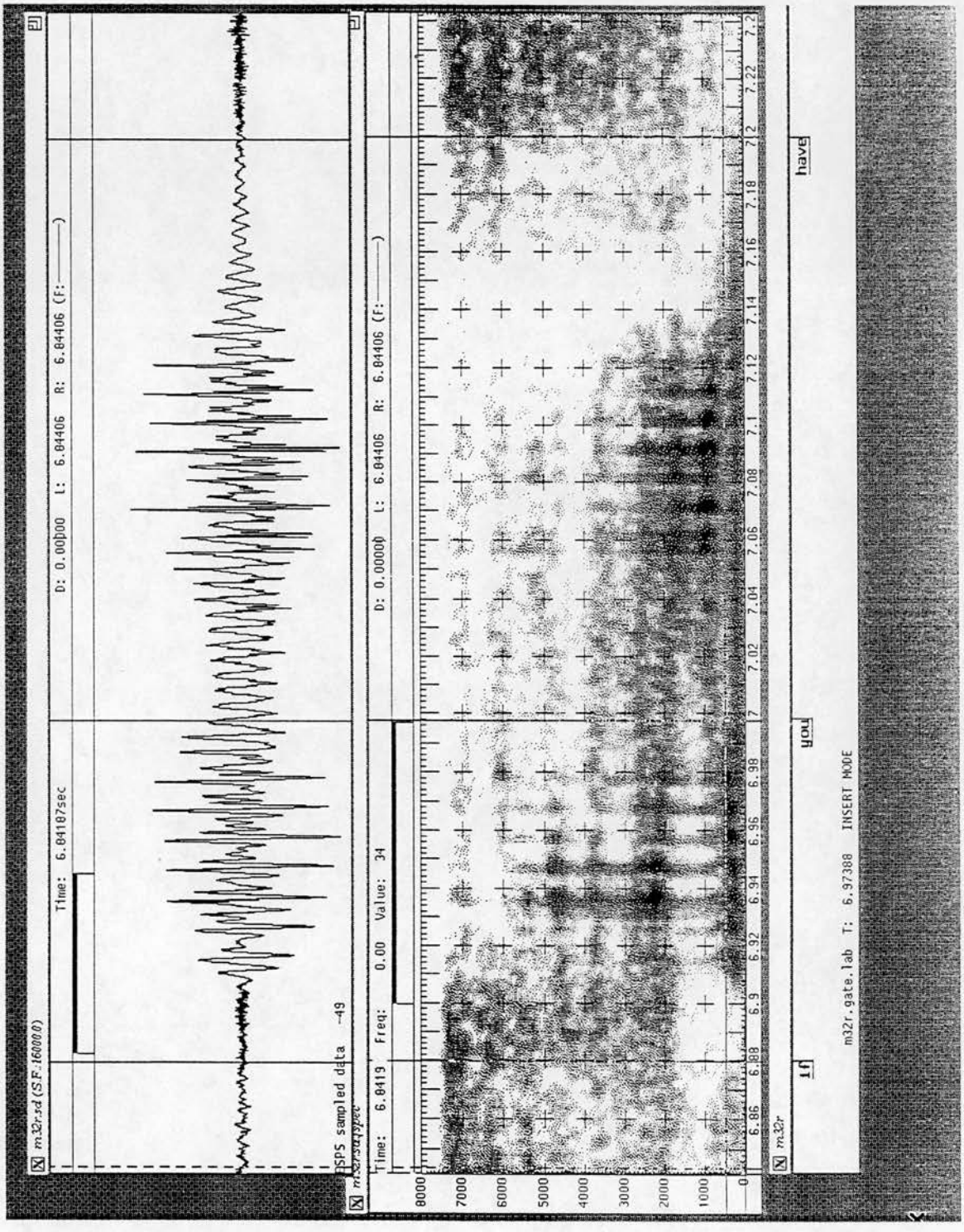

Figure 7.15. Acoustic Analysis: section of waveform and spectrogram and fluency judgements from Experiment 3 for "and if you have physiotherapy ...". Glottalisation in fluent speech, to compare with disfluent glottalisation in figure 7.14.

7.3.5 Juncture

The studies discussed in section 7.1.1, above, and our own analyses so far have looked for cues in pausing, F_0 values and glottalisation but have neglected another important phonological feature of continuous speech. Not only are words in fluent continuous speech usually *not* separated by silent pause into discrete units, but their boundaries are also usually linked or obscured by 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 signal into words on the basis of acoustic information: indeed, the problems of *word segmentation* are such that they provide fodder for a very active research area in psycholinguistics (Mehler *et al.*, 1981; Cutler & Norris, 1988; Sebastian, 1992; Quené, 1992; Cutler & Butterfield, 1992; Cutler *et al.*, 1992; Cutler & Mehler, 1993).

The degree to which linking between words takes place depends to some extent on sentence structure: phonological linking can be "blocked" by syntactic boundaries (Cooper & Paccia-Cooper, 1980; Egido & Cooper, 1980), which will often coincide with prosodic boundaries (Gee & Grosjean, 1983; Beach, 1991; Price *et al.*, 1991). Other factors, such as the placing of emphatic stress on one or other of the two words which share the boundary, may also block the link (Cooper *et al.*, 1982). But the likelihood of such blocking depends on speech style and speech rate (Cooper *et al.*, 1982; Lass, 1984): in the faster casual speech that occurs in spontaneous conversation, such boundaries are not respected by all types of linking, all the time. From the perceptual point of view, it is interesting to note that listeners are sensitive to the presence or absence of linking: Scott and Cutler (1984) show that palatalization and tapping at word boundaries in American English can help listeners to discriminate between syntactically ambiguous strings of words.

A large proportion of the disfluent stimuli we used in the experiments had no silence between the reparandum and the repair. The question that arises now is: does disfluency block phonological linking? If it does and if the reparandum does not end at a major syntactic or prosodic boundary, where linking might be blocked anyway, unexpected linking blocks may act as a cue for listeners. Indeed,

this may be the closest we will get to finding a signal to match Hindle's proposed "editing signal".

In order to address the question, waveforms and spectrograms of the disfluencies in our data set were examined for linking phenomena. In one case, where the reparandum ended at a major syntactic boundary, the disfluency was excluded from the analysis, as there was a possibility that no linking would occur in this place in fluent speech. Also excluded from the analysis were all disfluencies with silent pause greater than 50ms at the interruption ($N = 13$). Three items with fragment-final reparanda were also excluded from the analysis where it was impossible to hypothesise a fluent continuation, but the remaining disfluencies with fragments were included. A total of 13 utterances were examined. The boundaries between the reparandum offset and the repair onset were compared with hypothesised fluent boundaries between the same phonetic segments. So, for example, where the reparandum offset and repair onset consisted of *"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] with [j]. In the case of the three remaining items with reparanda ending in fragments, the hypothesised fluent continuation was assumed to be the continuation of the word that the fragment began. For the purpose of illustration, "fluent" versions of the disfluent utterances were recorded by the author, imitating the speech rate and intonation of the original utterances as closely as possible but making smooth links between the end of the reparandum and the onset of the repair.

Twelve of the 13 items examined had boundaries which differed from the hypothesised fluent boundary. The repair onset usually commenced as if it was being produced in isolation rather than as if it was preceded by a phonetic context. Repairs with voiced onsets had glottal stops at the onset:

- *"el-eligible"* contained a glottal stop (figure 7.16) where a fluent progression from [l] to [e] would show smooth formant transitions (figure 7.17);
- *"was was"* contained a glottal stop and a glottalised glide into the vowel, where a fluent link might contain assimilation (lip-rounding) in the fricative

(figures 7.18 and 7.19);

- *"valiency valency"* contained a weak glottal stop and prevoicing (fig. 7.20).

Other manifestations of non-linking involved the offset of the repair:

- in *"over the over the"*, the word "the" at the end of the reparandum was pronounced with schwa, as if a consonant was expected (figure 7.21 compare with the fluently produced figure 7.22));
- in *"ab-aberdeen"* the interrupted [b] is not audibly released.

The one item which did not have a clear "break" in the form of the absence of linking, *"you can easily WELL the FEES..."* (figure 7.23), had what we have described as a precipitous repair, with a large rise in intensity and sentence-initial prosody.

In Experiment 3, most disfluencies were detected within the first three gates, or 105ms, of the repair onset: it is possible that the lack of phonological linking was heard as an immediate cue to discontinuity. We noted earlier (section 7.3.4) that glottalisation often occurred in the reparandum but that subjects did not react immediately to it by giving more "disfluent" judgements, preferring to wait until the onset of the repair. It may be that glottalisation of vowel-final reparanda is one aspect of the phonological break we find at the interruption but that it is not until the onset of the repair that it can be confirmed that there is a break because listeners can not easily distinguish between epenthetic and interruption glottalisation.

7.4 Discussion

This study has examined pauses, pitch, glottalisation and word-boundary phenomena in the vicinity of disfluent interruption for the small set of data which formed the disfluent stimuli for the experiments described in Chapters 4 to 6 and compared the timing of cues with the responses in the disfluency judgement tasks. To conclude the chapter, we summarise the findings, compare them with previous studies and suggest future work.

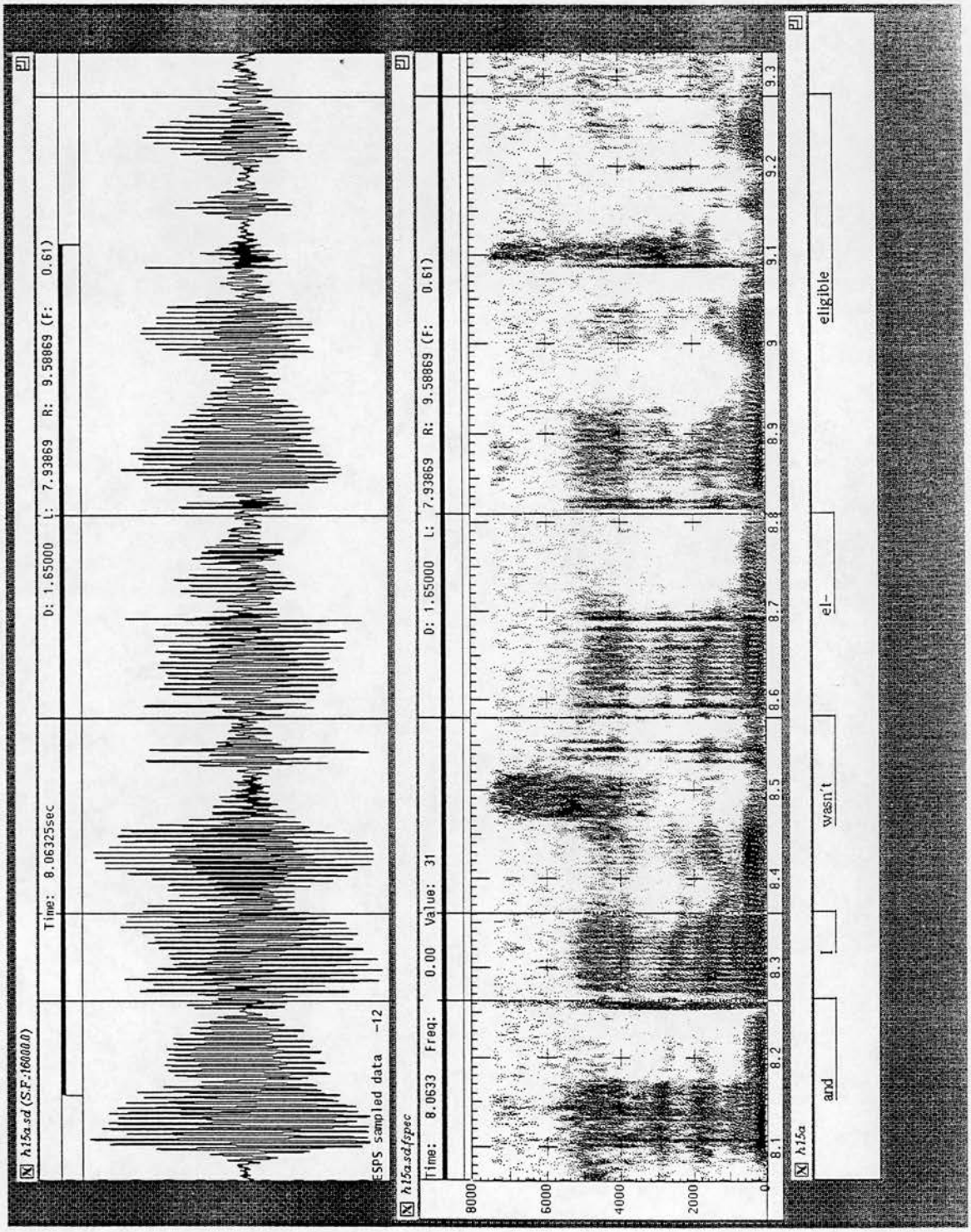

Figure 7.16. Acoustic Analysis: section of waveform and spectrogram from "and I wasn't **el-** — **eligible** for it". No break in voicing at interruption, but phonological break, as opposed to smooth transition. Compare with figure 7.17.

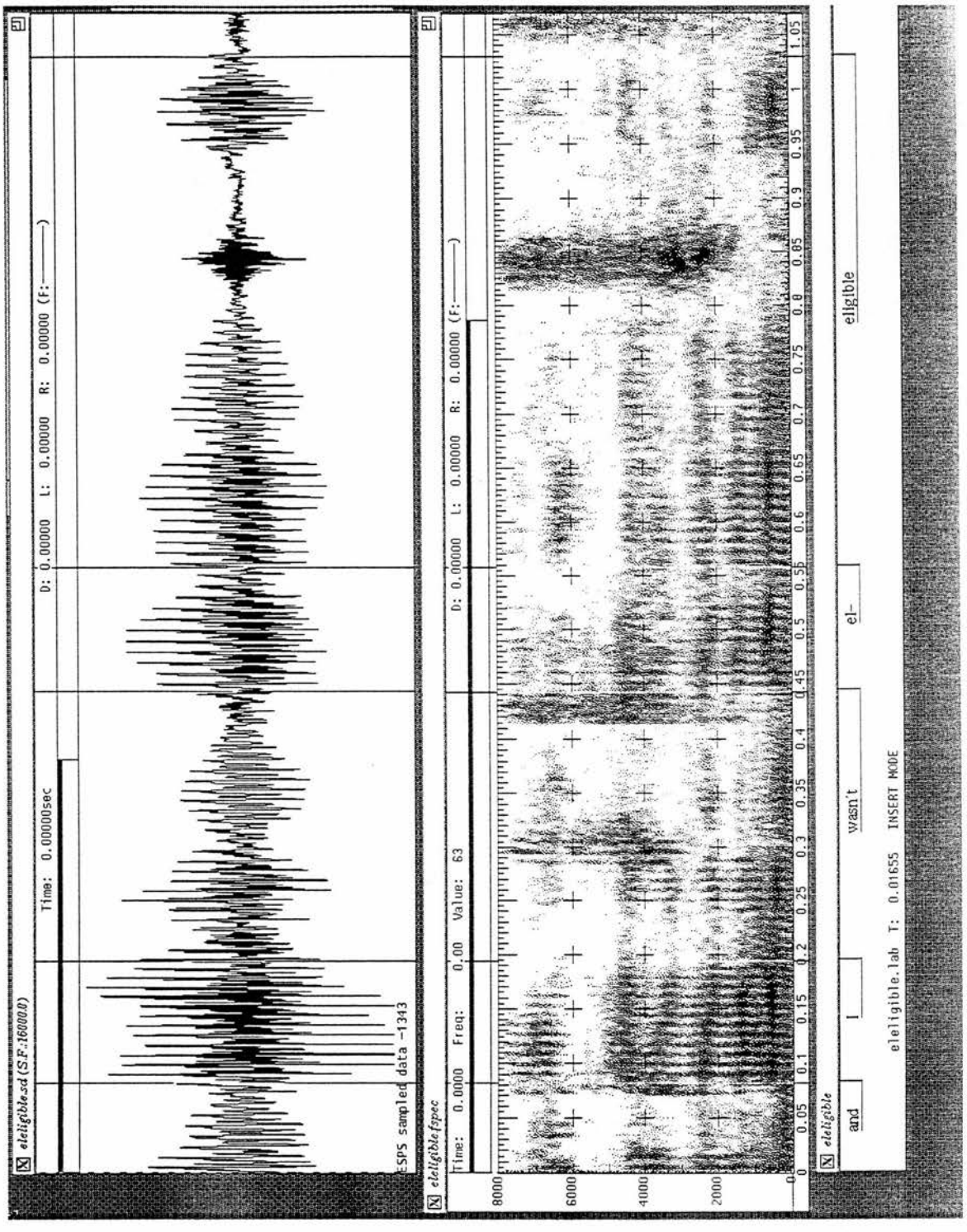

Figure 7.17. Acoustic Analysis: section of waveform and spectrogram from "and I wasn't **el-** — **eligible** for it". Fluently produced version to compare smooth transition [e hl - eh l] with break in figure 7.16.

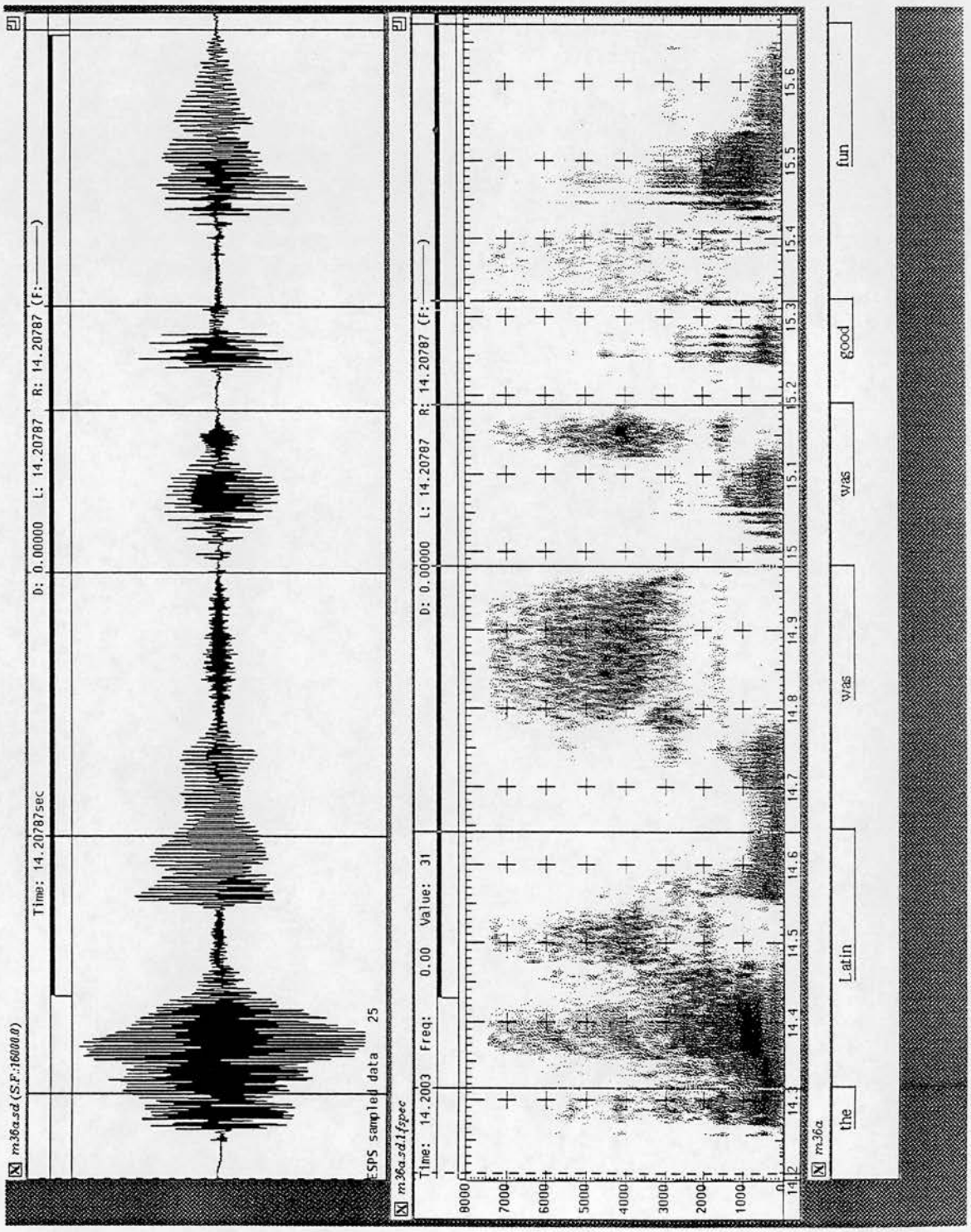

Figure 7.18. Acoustic Analysis: section of waveform and spectrogram from "the Latin **was** — was" good fun". Phonological break at interruption: No lip-rounding at end of fricative at offset of reparandum; glottalised onset to repair. Compare with figure 7.19.

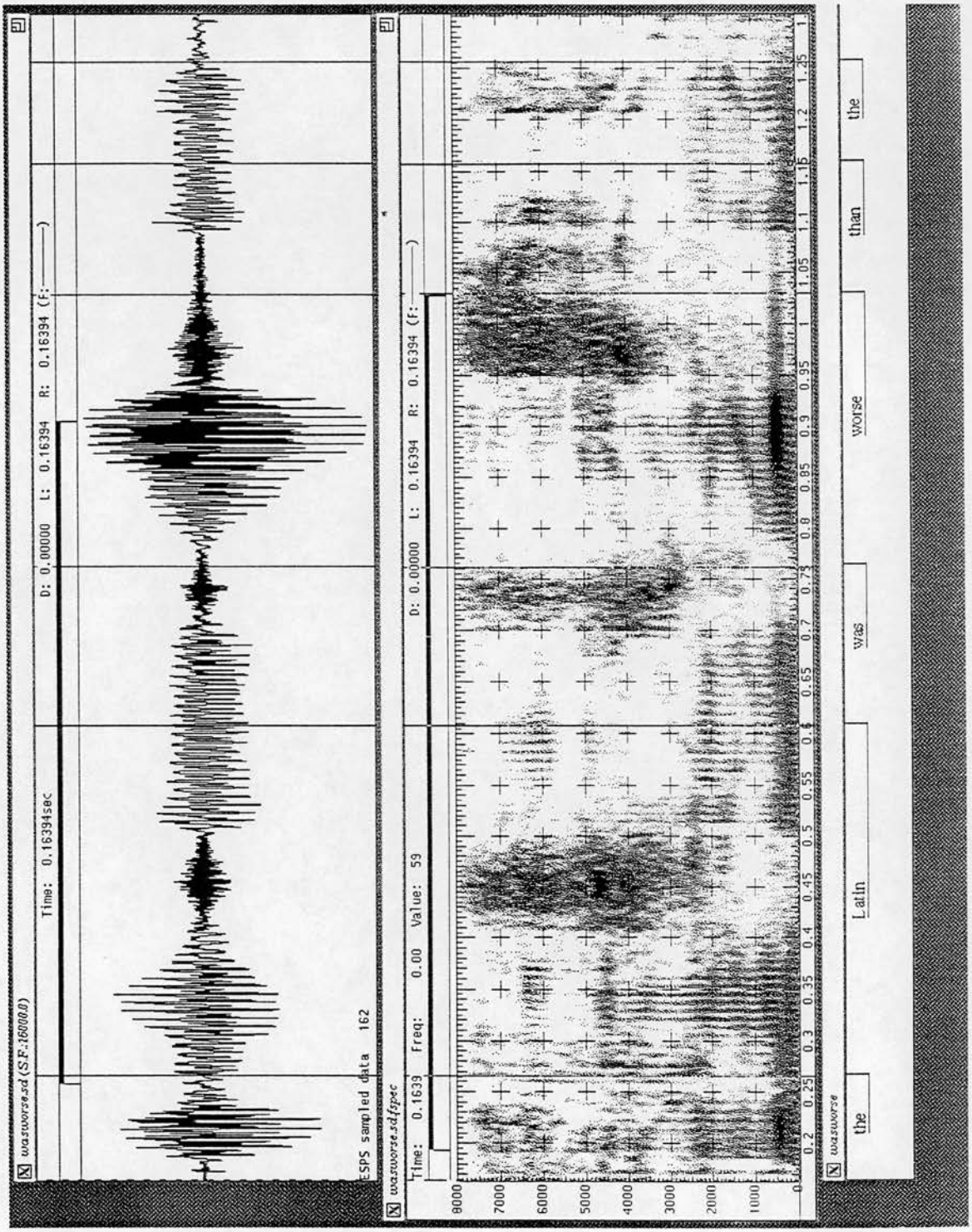

Figure 7.19. Acoustic Analysis: section of waveform and spectrogram from "the Latin **was — worse** than the Greek". Smooth [z] to [w] transition: liprounding (forward assimilation) at the end of the fricative; smooth voicing at the onset of [w]. Compare with break phonology of figure 7.18.

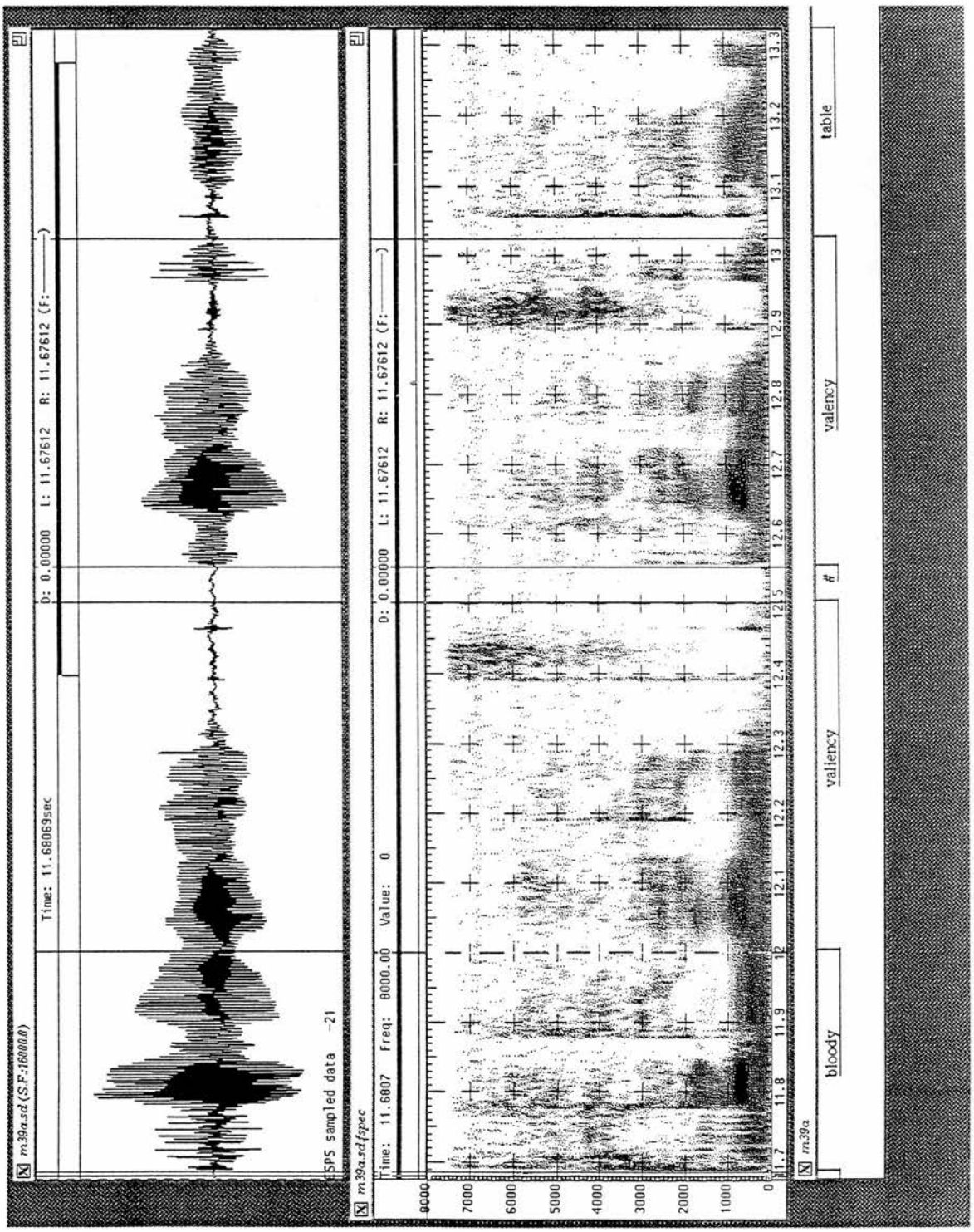

Figure 7.20. Acoustic Analysis: section of waveform and spectrogram from "and he'd always test me on the bloody **valiency** — **valency** table". Phonological break at interruption: onset of the repair has slight glottal stop and prevoicing.

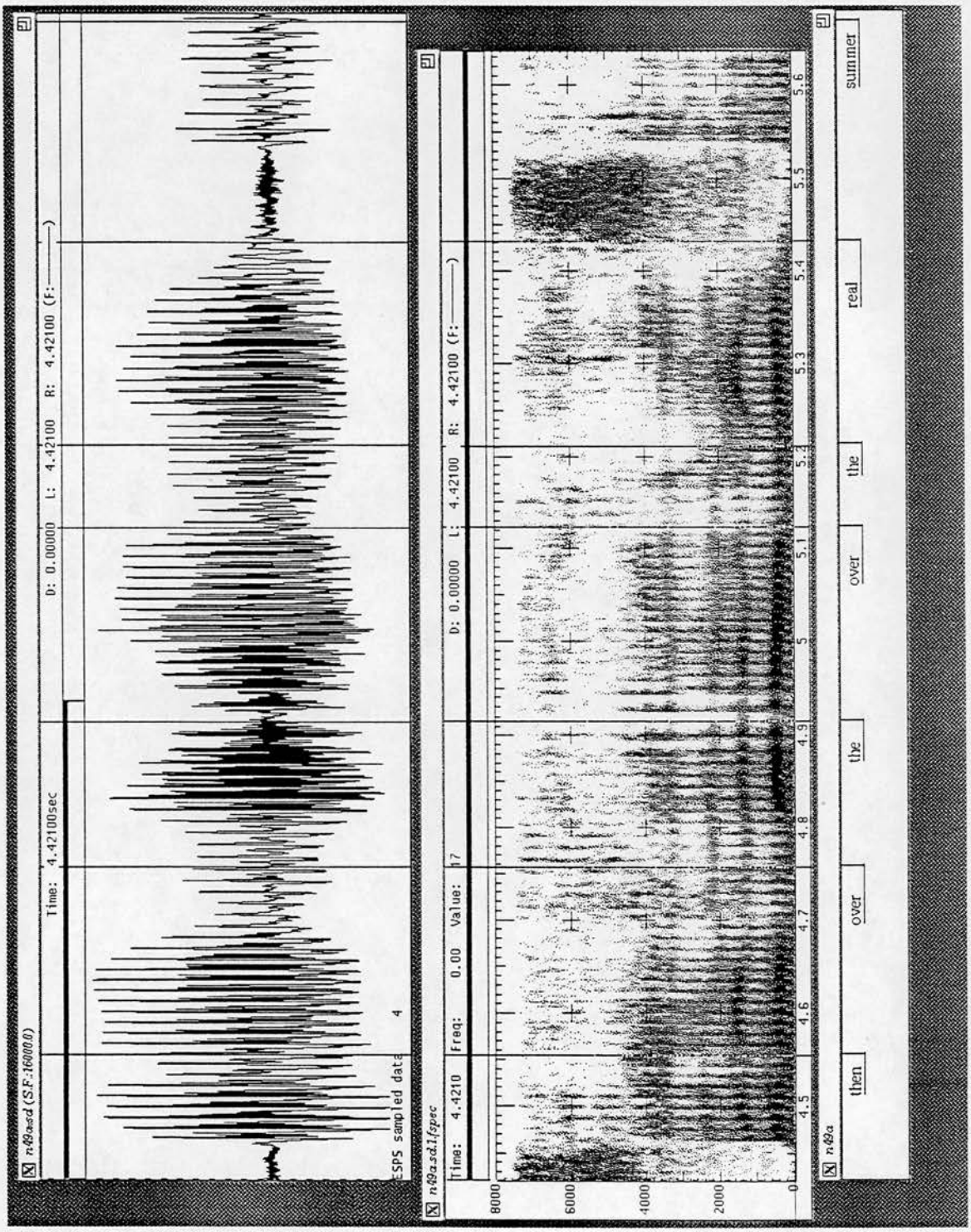

Figure 7.21. Acoustic Analysis: section of waveform and spectrogram from "then over **the** — **over** the real summer ...". Phonological break at interruption: "the" realised as [dh@] before the vowel-initial repair; glottal onset to repair. Compare with "fluent" version, figure 7.22.

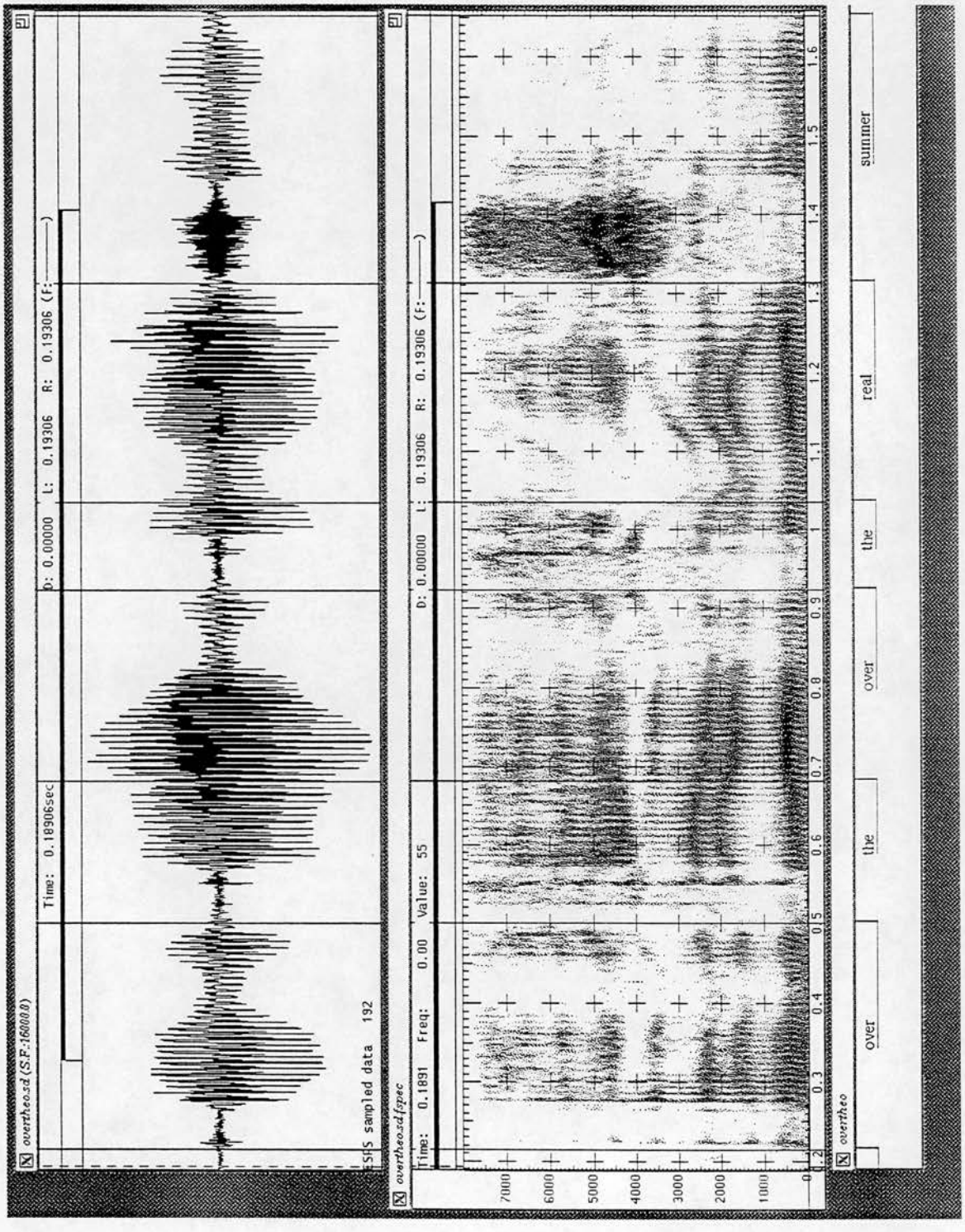

Figure 7.22. Acoustic Analysis: section of waveform and spectrogram from "then over **the** — **over** the real summer ...". Fluently produced version of the utterance in figure 7.21. "the" is realised as [dhi], with [j] linking to [ou] in "over"; smooth transition, with regular pitch pulses.

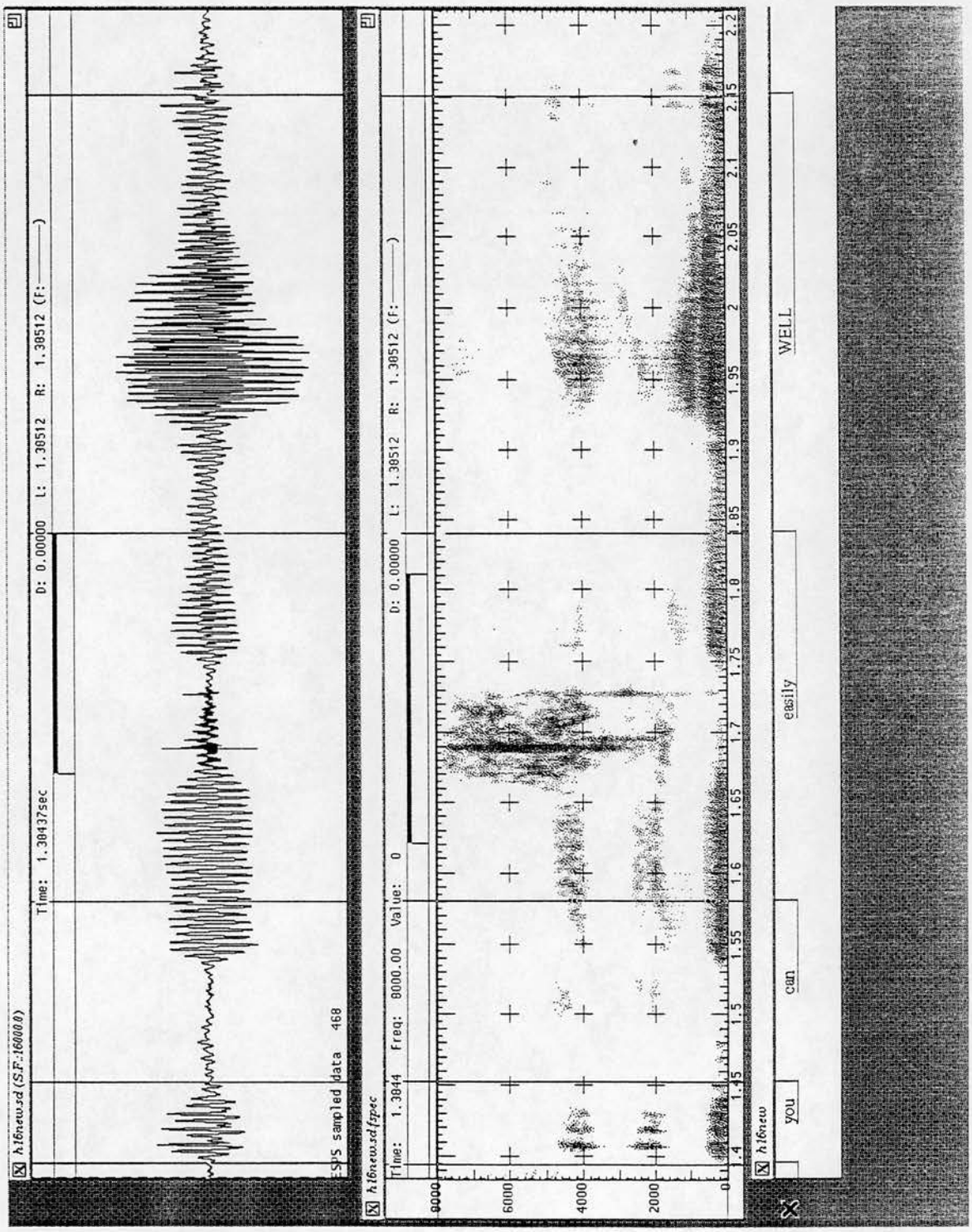

Figure 7.23. Acoustic Analysis: section of waveform and spectrogram from "since there's no fees you can **easily** — **WELL** the fees are ...". No evidence of phonological break, but low intensity of reparandum offset contrasts with higher intensity of repair onset.

Pauses accompanied the interruption point in about half of the stimuli examined. Some of these pauses contained inhalation or other vocal sounds. Where the pause was "clean" and longer than 105ms, it was possible to observe a gradual increase in the mean disfluency judgement scores (rising towards "disfluent") in the 35ms gating experiment (Experiment Three). A steeper rise was observed at the onset of the repair. Where the pause contained audible inhalation or other vocal sounds, subjects in the same experiment perceived this as a signal of discontinuity. We should not overestimate the importance of such "cues", however, since it may be that their prominence is exaggerated by the experimental technique and by the high quality of the digitised signal: in normal listening conditions, auditory information of this kind is likely to be obscured by ambient noise on the one hand and overlooked through more complex processing demands on the other. Whether or not these sounds have an effect themselves, there is still clearly an audible temporal break between the words at the interruption point where they were found. The presence of pause at a place where it is not predicted by syntactic or prosodic phrasing is in itself a sign of disfluency, but it is not necessarily a signal that repair is taking place, except where the "word" before the interruption can be clearly identified as a fragment: in most cases it is not until the onset of the repair itself that reliable detection can be achieved.

Other studies that have looked at pauses as possible cues to repair (e.g. Nakatani and Hirschberg, 1993a and 1993b, O'Shaughnessy 1992, 1993) have looked to the length of silence in order to distinguish disfluent from fluent pauses. Probably because of the general nature of the computational linguistics approach that they represent, the fact is overlooked that fluent pauses are usually accompanied by boundary prosody: the approach is bottom-up in the sense that the first problem is to detect silences in the speech signal and then to distinguish between fluent and disfluent silences and stop closures on the basis of the first available cue, duration, and does not take into account the features of syntactic and prosodic structure which make fluent pauses predictable and disfluent pauses perceptually more prominent (Butcher, 1981). Our study, like others, ignores the lengthening which accompanies, or, in many cases, constitutes perceived pause (e.g. Duez, 1993). Future work examining the rôle of pause in the detection of repair should take into account such lengthening.

In examining the pitch of the reparandum and repair, we first tested the *Reset Hypothesis*, which predicted that F_0 would fall less across the interruption in disfluent sentences than across control points in similar fluent sentences. Only partial support was found for the Reset Hypothesis, mainly in false starts rather than in repetitions, and the null hypothesis could not be rejected. Subsequent, more detailed analysis of the data suggested that the Reset Hypothesis, was too simplistic. Even where a repair commenced with a sentence-initial peak, its pitch was not necessarily higher than the previous peak, because the previous peak was often sentence-initial itself. Other repairs varied considerably in their relationship to the reparandum, making it difficult to posit any other F_0-related phenomena as cues. Examination of a considerably larger sample, taking into account repair type, reparandum length, and position in the syntactic and prosodic structure of the sentence of both the interruption and the repair might produce more revealing results.

Little is known about the rôle of rhythm in spontaneous speech. If English spontaneous speech does have a rhythmical pattern of stressed syllables and if listeners are as sensitive to perturbations in spontaneous speech as they appear to be to those in laboratory speech, then one feature of disfluent speech which could alert listeners to the presence of repair is the disruption of rhythm: in disfluent speech, the next stressed syllable may arrive earlier or later than the point predicted by the speech up to the interruption, depending on the nature and structural position of the repair. But such hypotheses can only be speculative until more evidence is available on the perception of rhythm in spontaneous speech. In addition, of course, stressed syllables are the loci for pitch peaks, so it is difficult to differentiate between effects of rhythm and effects of pitch in experiments with unadulterated spontaneous speech.

Discrete acoustic events which might signal disfluency were discussed in sections 7.3.4 and 7.3.5. Glottalisation was found in reparanda ending in voiced segments. This phenomenon was examined in a separate section from other word-boundary phenomena, mainly to make clear comparisons between our data and other studies which have looked at glottalisation as a possible cue. But, rather than giving glottalisation special status, we prefer to see it as another manifestation of the phonological break discussed in section 7.3.5. Phonological

breaks, as opposed to the *links* expected in fluent speech, were found in most repairs examined: assimilation or "linking phonemes" were not found. The restart phenomena at the repair onset, where word onsets were pronounced as if devoid of connecting left context, coincided with detection of disfluency in the results for Experiment Three. In one case, no phonological break was found, but the intensity and F_0 of the restart were strong indicators of a discontinuity. Previous studies have examined glottalisation as a signal of oncoming repair, but none have looked at it in the context of the general phenomenon of phonological linking. Other studies have found perceptual evidence that listeners are sensitive to the presence or absence of some types of linking when making judgements about phrase boundaries. Further study of the phenomenon from a human perceptual point of view in the context of disfluent speech would be worth pursuing.

It would be very satisfactory to conclude this chapter by describing a simple mechanism through which we can model the process of detecting disfluency in spontaneous speech. But the complexity of the phenomenon and the variety of forms which it can take make it impossible to construct such a mechanism from the small sample of data that we examined. What we have been able to do, however, is to show what cues listeners reacted to in the context of our experiments. The clearest cues were discontinuity in the form of pauses and the absence of phonological linking, where the syntactic and prosodic context predicted continuity in both. The impression of a break in the signal given by the lack of a phonological link at an interruption point is the closest we have found to the "abrupt cut-off" posited by Hindle (1983) as an editing signal, but it is only definable in terms of the distinction between what is expected in fluent speech and what occurs in repairs and cannot be defined as an isolatable and universal discrete signal. The experiments with low-pass filtered speech, where segmental information, including linking information, was obscured, but disfluencies still detected, suggest that higher level prosodic information is also of use in the processing of disfluent speech. Where a repair consisted of a complete restart of the sentence, the sentence-initial prosody was identified as the most obvious intonational cue to repair. Other intonational evidence may be useful, possibly in combination with rhythmic information, but further study of larger sets of data is needed.

Chapter 8

Conclusion

The work described in this thesis examined the perception of disfluencies in spontaneous speech with on-line processing experiments. The main aims were to find out how quickly disfluency could be detected and to identify what cues were used. To locate these cues, the best speech recogniser currently available was used: the human speech processing mechanism. It was found that disfluencies can be recognised very quickly, usually before the offset of the word after the interruption, even before the word itself has been recognised. Analysis of the points at which listeners were able to detect disfluency points to acoustic and prosodic features of the speech signal which can combine as cues.

The study was carried out in three major phases: data collection and analysis (Chapter 3); experimentation (Chapters 4, 5 and 6); acoustic and prosodic analysis (Chapter 7).

8.1 Data

A corpus of spontaneous speech was required to provide the raw materials for the experiments and analysis. At the beginning of the investigation, no suitable corpora were readily available, so the first task became to construct one. Six informal conversations of about 30 minutes each provided a sufficient supply of speech and disfluencies. The main aims of the identification of types of disfluency and the analysis of their relative frequencies were to assess how frequent the

phenomena are in normal conversational English – to demonstrate how great a problem they are for speech processors – and then to motivate the selection of stimuli to be used in the subsequent experiments.

The high frequency of disfluencies, which occurred every 9.4 words in our corpus, suggests that they should be of great interest for models of speech processing which have usually been based on either clean transcriptions or fluent, carefully prepared, read speech. As interest in speech processing moves more and more towards spontaneous speech, disfluencies are likely to become seen as normal, rather than as "ill-formed input".

On the basis of the relative frequency of types of disfluencies with repair identified in the corpus, a set of 30 disfluent stimuli and 30 fluent controls were selected. The same set of stimuli were used in all the experiments.

8.2 Detecting Disfluency

The general policy in approaching the question of *how soon* disfluencies could be recognised was first to look over a large portion of the speech signal and then, having established recognition points to a first approximation, to focus in on the crucial area, to identify more precise points. The gating technique allowed both these approaches to be employed under one experimental paradigm: for Experiments One and Two (Chapter 4), word-level gating was used; Experiment Three (Chapter 5) defined recognition points more accurately with 35ms gates.

Experiment One tested the hypothesis that an editing signal just before the onset of the repair could alert listeners to the oncoming discontinuity. The hypothesis was supported only in a minority of cases, where long pauses or clear mid-word interruptions prompted subjects to signal their detection of imminent disfluency. More judgements of "oncoming disfluency" were found for the word-gate following the interruption than for the previous gate. This suggested that subjects might in fact have been responding to actually perceived disfluency for the stimuli which contained pauses or fragments, rather than to cues of the nature of the proposed editing signal.

Experiment Two asked listeners to detect actual disfluency. It was found that listeners could reliably detect disfluency within the first word of the repair.

At this point in the investigation, it became interesting to speculate about the cues that listeners were using in identifying discontinuity so soon. Approaches from Computational Linguistics usually take a syntax-first approach, detecting repair sites, or potential repair sites on the basis of the occurrence of certain patterns of words or constituents: but that often entails the identification of a complete constituent rather than just the first word of a constituent, before a parse can fail. CL approaches also assume complete word recognition prior to syntactic processing: our on-line task demonstrated that listeners were sometimes able to detect disfluency even though they had not recognised all the words in the vicinity of the interruption. Nonetheless, word recognition generally was achieved relatively successfully in Experiments One and Two, so it was still possible that listeners used a "syntax first" strategy for most stimuli. Experiment Three looked more closely at the relationship between word recognition and disfluency detection.

Having established in Experiment Two that disfluency can be detected by the offset of the word after the disfluent interruption, Experiment Three was designed not only to find more precise recognition points, but to examine the relationship between points of disfluency recognition and points of word recognition. The 35ms gating technique allowed close comparison of these points. The results showed that listeners were able to detect disfluency very early in the word and , in the majority of cases, *even before the word was recognised*. A control experiment showed that the dual task of word recognition and disfluency detection had *not* caused a delay in word recognition which could have challenged this conclusion.

If disfluencies are detectable without the syntactic information provided via lexical access, then acoustic and prosodic information is the most likely source of cues. The last two experiments (Chapter 6) investigated the rôle of prosodic information in the detection of discontinuity, by presenting stimuli which had been low-pass filtered to remove all segmental information from the signal, leaving audible only intonation and relative amplitude.

In Experiment Four, the whole utterance was presented, low-pass filtered from the point of interruption, so that prosodic expectations based on the whole signal up to that point would not be diluted. Listeners were clearly able to distinguish between fluent and disfluent stimuli presented in this way. Prosodic information

(intonation, rhythm, pause and duration) was assumed to be responsible for the result.

Experiment Five combined 35ms gating and low-pass filtering, to assess the value of prosodic information as early as the first word of the continuation. The same gating method as was used in Experiment Three meant that subjects heard the two words on either side of the interruption presented in 35ms increments, but in this experiment the two words were also low-pass filtered. The results showed that even in speech degraded by low-pass filtering, listeners were able to detect disfluency early in the signal, by the offset of the first word of the continuation, by using prosodic information.

8.3 Acoustic Analysis

The experimental results, and particularly the responses in the 35ms gating experiments prompted a close examination of the speech signal in the vicinity of the interruptions in the stimuli for acoustic and prosodic cues.

The great variety of possible combinations of features at disfluent interruptions make the task of finding simple universal cues difficult and probably unrealistic. The study in Chapter 7 looks at several different possible sources of cues to disfluency. The two major cues identified are pause and the absence of phonological linking at the interruption point. Relative pause lengths have been discussed as cues by other authors (Chapter 2), comparing disfluent with fluent pauses, but not from the viewpoint of on-line processing. We made the observation that where there is no pause, the interruption site is characterised by the absence of a phonological link between the word before the interruption and the word after it, where in fluent speech adjacent words *are* usually linked. This is a likely cue which has eluded other recent studies. Glottalisation around the interruption, which *has* been posited as a cue (Bear *et al.*, 1992; Shriberg *et al.*, 1992; Nakatani & Hirschberg, 1993a; Nakatani & Hirschberg, 1993b) is probably often a manifestation of broken linking, when it is not a pausal lengthening feature. Pitch differences across the interruption are difficult to generalise over. However, if we assume that listeners have prosodic expectations at any given point in the processing of an utterance, then it may be possible to suggest cases where the

expectation has failed because the direction of the pitch pattern changes in a way that is not compatible with its left context. This is a possible explanation for the ability that subjects in Experiment Three displayed to detect disfluency in whole utterances where the continuation was low-pass filtered. But the amount of data in our study was too small for any such suggestions to be tested.

8.4 Word Recognition

The gating technique allowed word recognition to be tested at the same time as disfluency detection, in all of Experiments One, Two and Three. The hypotheses regarding word recognition that were of greatest interest apart from the relationship between disfluency detection and lexical access concerned the recognition of words in the immediate vicinity of the interruption.

The word before the interruption had a left context as informative as the controls and was therefore expected to be recognised immediately as frequently as similar words in the controls. If it was not recognised immediately, it was expected not to be recognised at all, as it lacked the right context necessary for its successful recognition, whereas a similar word in a fluent right context was expected to be recognised late, rather than missed. Both these expectations were confirmed.

The word following the disfluent interruption, on the other hand, lacked a cohesive *left* context, which was present for its fluent controls. For this reason it was expected that the word would not be recognised immediately as easily as words in a similar serial position in the fluent control stimuli. This expectation was also confirmed by the results of all three experiments.

8.5 Discussion

This Chapter began with a description of the aims of the thesis. The sections that followed described to what extent the aims were achieved. The first question was *"how soon can disfluency be detected?"*. Experiments One to Three show that disfluency can be detected within the first word after interruption and usually

before that word is recognised. The second question was *"what cues can the listener use to detect disfluency?"*. Experiment Three shows that information in the speech signal available before the lexical information which allows access to syntax can be used. Experiments Four and Five showed that prosodic information may have a key rôle; the acoustic and prosodic analysis in Chapter 7 suggests cues in pauses and the lack of phonological links between words at a disfluent interruption.

This study constitutes a first step in the investigation of the on-line processing of normal disfluent speech. In the absence of previous work in the field, a small sample of different types of disfluencies was taken. In future work it would be of interest to use larger samples of data, divided into different types, both for experimentation and for analysis of the signal. It is likely, for example, that repairs which contain full or partial repetitions will have different implications for processing than repairs which contain completely new material; repairs involving sentence restarts might be expected to have different prosodic features from other repairs; pronunciation corrections may be easier to anticipate than other types. On a different level of analysis, the observation that disfluencies are easily missed in normal circumstances may suggest that the human speech processing mechanism has the ability to filter out disfluency without even coming across the potential problems to the processor inherent in discontinuity. There are many other hypotheses to entertain. The possible avenues for future research in this field seem boundless.

Appendix A

Appendix A: Materials

The experimental materials used in all experiments are listed here in blocks by speaker (G,H,J,M,N,P) and reference number (1-5). Codes A-C give the stimulus type, where:

A = Spontaneous Disfluent Stimuli;
B = Rehearsed "Disfluent" Stimuli;
C = Spontaneous Fluent Stimuli and Rehearsed Fluent Stimuli.

Parentheses mark words on either side of interruption or matched point in controls. Stimuli H5 and N5 were omitted from Experiment 3.

G1A no what [we we] do is we look at statistics
G1B no what [we do] is we look at statistics
G1C so what [one does] stops taking ...

G2A well in [Edinburgh no] I think in Edinburgh it's been quite ...
G2B well in [Edinburgh it's] been quite active
G2C so for a few [days you] can be quite upset

G3A they [sent a] lot of their youngsters would go off ...
G3B they [sent a] lot of their youngsters
G3C they [send their] employees to us

G4A they certainly are [w@- to] alcohol
G4B they certainly are [to alcohol]
G4C who are in trouble [with alcohol]

G5A it's quite obvious [he's he's] on something
G5B it's quite obvious [he's on] something
G5C we know that it's [not going] to be ...

H1A and I wasn't [el- eligible] for it
H1B and I wasn't [eligible for] it
H1C I wasn't [good enough] for them

H2A I didn't [like like] them by any means
H2B I didn't [like them] by any means
H2C they weren't at all [vocal until] very recently

H3A and then [you if] you wanted to graduate
H3B and then [if you] wanted to graduate
H3C and if [you wanted] to go on in English ...

H4A but in [some some] English Universities
H4B but in [some English] Universities
H4C d'you mean for [the degree] structure and so on

H5A and since there's no fees you can [easily well] the fees ...
H5B and since there's no fees you can [easily just] work ...
H5C 'cos they're not really very difficult you can [easily just] cram in ...

J1A cos I-I think it's a much [more I] find it ...
J1B cos I-I think it's a much [more permanent] kind of ...
J1C I think you were living [in Lussielaw] Road then weren't ...

J2A I don't know what [the I] don't know what the outcome will be
J2B I don't know what [the outcome] will be
J2C I don't know I don't know what [the situation] will be like

J3A it's especially a problem in [ab- Aberdeen] apparently
J3B it's especially a problem in [Aberdeen] apparently
J3C well it's a beautiful place to [live I] would have ...

J4A but they've thrown away [that that] trump card
J4B but they've thrown away [that trump] card
J4C oh the SNP has got [a very] developed set ...

J5A I mean a [normal uh] if you went to a teacher ...
J5B I mean [if you] went to a teacher ...
J5C I mean any [country has] to have a dominant ...

M1A there [wasn't there] wasn't a great deal of choice
M1B there [wasn't a] great deal of choice
M1C there [were various] parts to the art A-level

M2A he'd always test me on the bloody [valiency valency] table
M2B he'd always test me on the bloody [valency table]
M2C and you'd always stop and have a [game on] the way

M3A and if [you it] it just it just sometimes gets ...
M3B and if [you had] physiotherapy it's go down a bit ...
M3C because if [you stay] to the back of him ...

M4A both of [your both] styles adapt
M4B both of [your styles] adapt
M4C like one of [my friends] wanted to do music

M5A the latin [was was] good fun
M5B the latin [was good] fun
M5C and that [was quite] good

N1A one of the things I thought the [psych- there's] a psychologists' ...
N1B one of the things I thought the [psychologists could] do ...
N1C um I don't know I think it's [just a] check actually

N2A then over [the over] the real summer ...
N2B then over [the real] summer I'm really ...
N2C but um [when I] and the people in years before me...

N3A it's word meaning [vei- um] sort of very vaguely ...
N3B it's word meaning [very vaguely] word meaning
N3C it's a bit [daunting actually]

N4A it's so much easier [to to] put it in the corner
N4B it's so much easier [to put] it in the corner
N4C it would be cheaper [to hire] a car between us

N5A cos on [Sat- Sunday] Sunday I'm going to America
N5B cos on [Sunday I'm] going to America
N5C well [Christmas I] find rather special

P1A I think what'll happen is that [it'll the] general movement ...
P1B I think what'll happen is that [it'll have] a pull
P1C the problem I always worry about is that [you have] situations ...

P2A I don't know if [it how] true it is
P2B I don't know if [it's true] ...
P2C I don't know if [the right] in Britain ...

P3A the idea is [apparent-? is] apparently ...
P3B the idea is [apparently quite] successful
P3C but they're [probably quite] valid

P4A I've never understood [how how] you can be into psychological things ...
P4B I've never understood [how you] can be into psychological things ...
P4C you get situations [where they] watn separate schools

P5A um they actually [kh- commit] the cardinal sin...
P5B um they actually [commit the] cardinal sin ...
P5C the group [becomes less] and less distinct

Appendix B

Appendix B: Publications

The following papers were published by me during the course of the preparation of this thesis.

LICKLEY, R.J., & BARD, E.G. 1992 (October). Processing Disfluent Speech: Recognising Disfluency Before Lexical Access. *Pages 935–938 of: Proceedings of The ICSLP.*

LICKLEY, R.J., R.C., SHILLCOCK, & BARD, E.G. 1991a (September). Processing Disfluent Speech: How and When are Disfluencies Found? *Pages 1499–1502 of: Proceedings of Eurospeech 91*, vol. 3. 2nd European Conference on Speech Communication and Technology, Genova, Italy.

LICKLEY, R.J., BARD, E.G., & R.C., SHILLCOCK. 1991b (August). Understanding Disfluent Speech: is there an Editing Signal? *Pages 98–101 of: Proceedings of the ICPhS*, vol. 4. International Congress of Phonetic Sciences, Aix-en-Provence, France.

SHRIBERG, E.E., & LICKLEY, R.J. 1992a (October 12-16). Intonation of clause-internal filled pauses. *Pages 991–994 of: Proceedings of the International Conference on Spoken Language Processing*, vol. 2.

SHRIBERG, E.E., & LICKLEY, R.J. 1992b. The relationship of filled-pause F0 to prosodic context. *Pages 201–209 of: Proceedings of the IRCS Workshop on Prosody in Natural Speech, Technical Report IRCS-92-37.*

SHRIBERG, E.E., & LICKLEY, R.J. 1993. Intonation of clause-internal filled pauses. *Phonetica*, **50**, 172–179. **[In Press]**

PAPER 1

LICKLEY, R.J., & BARD, E.G. **1992 (October). Processing Disfluent Speech: Recognising Disfluency Before Lexical Access.** *Pages 935–938 of: Proceedings of The ICSLP.*

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.

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 dis-

fluencies 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.

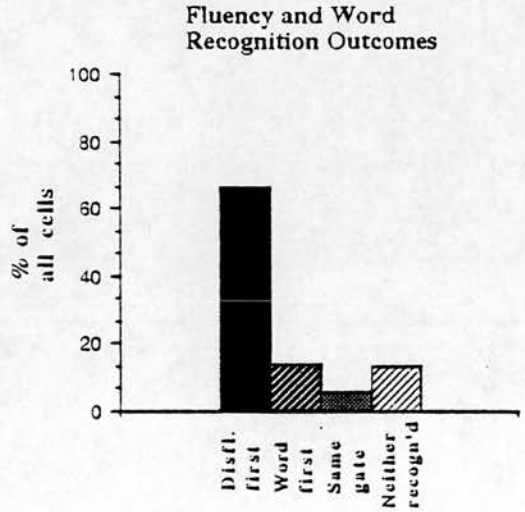

Figure 1.

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).

PAPER 2

LICKLEY, R.J., BARD, E.G., & R.C., SHILLCOCK. **1991b (August). Understanding Disfluent Speech: is there an Editing Signal?** *Pages 98–101 of: Proceedings of the ICPhS*, **vol. 4. International Congress of Phonetic Sciences, Aix-en-Provence, France.**

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 of 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 too 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.*

PAPER 3

LICKLEY, R.J., R.C., SHILLCOCK, & BARD, E.G. **1991a (September). Processing Disfluent Speech: How and When are Disfluencies Found?** *Pages 1499–1502 of: Proceedings of Eurospeech 91*, **vol. 3. 2nd European Conference on Speech Communication and Technology, Genova, Italy.**

PROCESSING DISFLUENT SPEECH: HOW AND WHEN ARE DISFLUENCIES FOUND?

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

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

Abstract

Disfluency in spontaneous speech presents problems for both psycholinguistic and computational models of speech understanding. However, it is not clear that the human speech processing mechanism is greatly disrupted by the presence of disfluency. This paper presents the results of three experiments on the perception of spontaneous speech: two gating experiments show that disfluency can usually be recognised by the end of the word following a disfluent interruption, while listeners' ability to recognize words is not greatly affected by the presence of disfluency; the third experiment, using low-pass filtered speech, suggests that prosodic information may have a key role in aiding the processing of disfluent speech.

Keywords: speech perception; word recognition; spontaneous speech; disfluency; gating experiments; low-pass filtered speech.

1 Introduction

An important feature of spontaneous speech which is absent in printed text or read speech is the frequent occurrence of hesitation phenomena or disfluencies. For the purposes of 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 often surprising to see how a correct transcription of spontaneous speech is peppered with repetitions, false starts and disfluent pauses, since we appear to notice very few of them as they occur and to recognize speech despite them. This suggests that the human speech processing mechanism is not greatly disrupted by the presence of disfluency, though machine speech recognizers and parsers, which are modelled on carefully prepared, scripted materials might be very vulnerable to the failure of normally disfluent speech to conform to their models. Hindle [1] proposes a parser which requires external and dependable indications of the existence of a disfluency in order to deal appropriately with what otherwise might be an unparsable string. Because very little work has been done on the perception of disfluent speech, it is not known whether there are any clear indications of disfluency at or around the point where the flow of speech is interrupted and if there are such indications, what form they take.

This paper presents results from three experiments designed to look at how and when disfluency can be recognised. 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 two experiments involve two tasks: disfluency recognition and word recognition. Results of the disfluency recognition task are described by Lickley, Shillcock and Bard [2] and will be summarised in Section 2 below. The results of the word recognition task are also given. Together they show that a disfluency can rarely be recognised before the onset of a disfluent interruption but can usually be recognised by the offset of the word following this. The presence of disfluency has only a slight effect on listeners' ability to recognise words. The third experiment makes use of low-pass filtering to remove segmental information from the speech signal and shows that listeners are able to judge the fluency of spontaneously produced utterances on the basis of prosodic and temporal phenomena, even when they cannot recognise any of the words involved.

2 Gating Experiments

2.1 Introduction

These experiments were designed to find a recognition point for disfluency to a first approximation. Since one of the tasks in the experiments was to recognise the words in each utterance, it was also possible to test for any effect of the presence of disfluency on the subjects' ability to recognise words.

2.2 Materials

The utterances used in these experiments were selected from a corpus of spontaneous speech digitally recorded in a studio. Thirty disfluent utterances were chosen as representative of the types of disfluency present in the whole corpus. Fluent control

independent signal of disfluency might prove more efficient than detecting disfluency on the basis of continued failure to parse or recognize a disfluent string. As a preliminary investigation into this possibility, an experiment was devised using low-pass filtered speech to test the hypothesis that listeners are able to detect disfluency using prosodic information even when lexical information is lacking.

3.2 Materials

The 30 disfluent and 30 spontaneous fluent utterances, matched for structure, length and prosody, which were used in the first two experiments, were also used here.

Each disfluent utterance was low-pass filtered from the interruption point (ie the point at which the fluent continuation begins) to the end of the utterance. The fluent utterances were also low-pass filtered from the equivalent point. The level of the filter was set individually for the six different speakers whose utterances made up the materials so that, while rhythmic and intonational information was preserved, it was impossible to hear any segmental information.

The utterances were blocked by speaker and presented in random order within each block. Each utterance was presented three times, the first presentation being preceded by about 10 seconds of the conversation prior to the test utterance.

3.3 Subjects and Procedure

Twelve students of the University of Edinburgh served as subjects. All were native speakers of English and all reported having normal hearing.

The experiment was run in two sessions of 20 minutes.

The subjects were asked to listen carefully to the utterances and to make a judgement on a scale of 1-5 as to whether they thought the filtered speech continued fluently or disfluently from the unfiltered introduction. A score of 1 would signify that the subject thought that the continuation was fluent, 5 disfluent.

At each of the three presentations of an item, subjects were asked to make a new judgement.

3.4 Results

Over the 360 paired judgements, the overall mean score for disfluent items was 3.36 and for fluent items, 1.90. The difference was highly significant (W=4519, p<.0001). The score for disfluent items was greater than for fluent items in 229 cases of 271 score-pairs that had non-zero differences (Sign test at p<.0001). The mean score for each disfluent item was significantly higher than that for the fluent control in 28 out of the 30 cases. Similar results were found for the first and second presentations, with the same levels of significance.

Positive correlations were found between the mean fluency judgement for each item and both the length of pause at the interruption point (n=30, R=0.529, p=.002) and the length in syllables of the reparandum (ie. the speech that is to be ignored as a result of the disfluency) (n=30, R=0.382, p=.036).

3.5 Discussion

The experiment showed clearly that listeners were able to judge disfluency by hearing prosodic information alone from the point of interruption. Correlation tests showed that this ability may be related to the length of the pause that often occurs at a disfluent interruption and to the size of the disruption to the fluency in terms of the number of syllables of "extra speech" that occur. It is so far unclear whether a disruption in the pitch contour can also contribute to listeners' perception of disfluency.

4 Conclusions and Future Work

From the results of the experiments reported here, we may conclude that it is possible for listeners to detect disfluency very soon after they occur and usually within one word of the point of interruption. A more precise location of a recognition point for disfluencies is under investigation.

The results of the experiment with filtered speech support the view that some form of prosodic information plays a vital role in marking the presence of disfluency: if listeners make use of rhythmic expectations in processing speech ([5], [6], [7]) and are aware of pitch continuity in fluent speech ([3], [4]) then it is likely that listeners are alerted to the presence of disfluency when these expectations fail or when the pitch contour becomes discontinuous.

Experiments are currently in progress to establish more precisely how quickly disfluency may be recognised. These experiments use much shorter gates then the word-length gates used in the experiments described here: using such a method, it should also be possible to show whether or not it is possible for listeners to make sound fluency judgements before having recognised the word they are hearing.

Further experiments with low-pass filtered speech are also in progress: these experiments also use the gating method to look more closely at the point of recognition of disfluency in filtered speech.

References

[1] Hindle, D., 1983. *Deterministic Parsing of Syntactic Non-Fluencies.* In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pp123-128. ASL.

[2] Lickley, R.J., Shillcock, R.C. and Bard, E.G., 1991. *Understanding Disfluent Speech: is there an Editing Signal?* Proceedings of the XIIth ICPhS, Aix-en-Provence, August 1991.

[3] Darwin, C.J., 1975. *On the Dynamic Use of Prosody in Speech Perception.* In Cohen, A. and Nooteboom, S.G. (eds), *Structure and Process in Speech Perception.* Springer-Verlag: Berlin.

[4] Nooteboom, S.G., Brokx, P.L. and De Rooij, J.J., 1978. *Contributions of Prosody to Speech Perception.* In Levelt,

utterances were also selected, matching the disfluent utterances for structure, length and prosody as far as possible. Two further sets of control utterances were produced: these were rehearsed versions of the first two sets (without disfluency), produced by a method described in [1].

All the speech material used in these experiments was sampled at 20kHz through a 8kHz filter.

Each utterance was presented with about 10 seconds of the conversation prior to it, providing some discourse orientation.

Each test utterance was *gated* at word boundaries so that the first presentation of each utterance ran from its onset to the moment prior to the onset of the second word ('*because* ... '), the second from its onset to the moment prior to the onset of the third word ('*because although* ... ') and so on ('*because although the* ...', '*because although the bell* ...', ...).

2.3 Subjects and Procedure

Twenty students and staff members of the University of Edinburgh served as subjects in each experiment, 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.

Both experiments were run in two sessions of approximately 45 minutes.

There were two tasks in both experiments: word recognition and disfluency recognition. For the *word recognition task*, subjects were asked to write down what they thought the latest word they had heard was as well as making any amendments that were required to previous judgements. The difference between the experiments lay in the *disfluency recognition task*: in the first experiment, subjects were asked to give a judgement as to whether they thought the utterance was fluent or disfluent at the current word gate; in the second experiment, subjects were asked whether they thought the utterance would continue fluently or disfluently. In both experiments the fluency judgement was given on a 1-5 scale, 1 indicating a judgement of "fluent" and 5, "disfluent".

2.4 Results

2.4.1 Disfluency Judgements

Subjects were able to detect disfluency in the word following the interruption with considerable certainty in the vast majority (28 out of 30) of cases. The differences between the judgements in the disfluent cases and in the controls was highly significant (Friedman statistic by subjects = 38.2, $df = 3, p < .001$; by materials = 50.91, $df = 3$, $p < .001$).

Where the fluency judgement suggested that subjects were less certain that disfluency was present, the judgement scores were still significantly higher than in the fluent controls ($p<.025$)

Subjects could, however, predict an oncoming disfluency in only a minority (12) of the 30 test items. All the detectable cases contained words which were either interrupted suddenly or contained or were followed by noticeable pauses. Elsewhere subjects were unable to predict the oncoming disfluency.

2.4.2 Word Recognition

For the purposes of this study, word recognition outcomes were classed as "right" where the word was recognised correctly on its first presentation, and "wrong" where the word was not recognised at all or recognised at a later presentation of the same test item.

To determine whether there was any effect on word recognition performance of the presence of disfluency in an utterance, the word recognition outcomes for the word prior to and the word following disfluent interruption were compared to the outcomes for the equivalent words in the spontaneous and rehearsed control utterances.

For the first experiment, there was no significant difference between word recognition performances in the vicinity of a disfluency and those at control points in fluent utterances.

For the second experiment the presence of disfluency did reduce subjects' ability to recognise the adjoining words significantly (all four chi-square tests produced significant results $p<.05$), but not greatly: around 20% of recognitions failed at these points in disfluent utterances, while the overall recognition failure rate for spontaneous fluent utterances is 15.3%.

There is no obvious reason why the results should differ for the two experiments: the materials and listening conditions were the same in both cases, the only difference being in the disfluency judgement task.

2.5 Discussion

The gating experiments showed that information in the first word following the disfluency is usually sufficient to tell the listener that a disfluency has occurred. This may be taken as evidence that there is not usually a discrete phonetically identifiable editing signal immediately prior to the onset of the fluent continuation (as proposed by Hindle [1]) but that some feature of the following word informs the listener that fluency has broken down.

One possible cue to disfluency that may be contained in the continuation is in its rhythmical and intonational properties. To investigate this possibility, the following experiment, using low-pass filtered speech, was performed.

3 Filtered Speech Experiment

3.1 Introduction

Prosodic information has been shown to be useful to listeners in understanding fluent speech. Darwin [3] showed that prosodic continuity could help a listener attend to a particular speaker when there was potential interference from other speech. Nooteboom, Brokx and de Rooij [4] cite evidence that pitch continuity helps listeners to perceive speech as belonging to a single auditory unit. Martin ([5], [6]) and Buxton [7] showed that rhythmic expectancy may be important in helping listeners to understand speech.

It thus seems likely that prosodic information is used in ways which would be helpful in processing *disfluent* speech. An

WJ.M. and Flores D'Arcais, G.B. (eds), *Studies in the Perception of Language*. John Wiley and Sons.

[5] Martin, J.G., 1972. *Rhythmic (Hierarchical) Versus Serial Structure in Speech and Other Behaviour.* In Psychological Review, Vol 79 No 6 pp487-509.

[6] Martin, J.G., 1975. *Rhythmic Expectancy in Continuous Speech Perception.* In Communication and Cybernetics, Vol 11.

[7] Buxton, H., 1983. *Temporal Predictability in the Perception of English Speech* In Cutler, A. and Ladd, D.R. (eds), *Prosody: Models and Measurements* Springer-Verlag: Berlin.

PAPER 4

SHRIBERG, E.E., & LICKLEY, R.J. **1992a (October 12-16). Intonation of clause-internal filled pauses.** *Pages 991–994 of: Proceedings of the International Conference on Spoken Language Processing*, **vol. 2.**

ICSLP '92

INTONATION OF CLAUSE-INTERNAL FILLED PAUSES

Elizabeth E. Shriberg

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025 USA

Robin J. Lickley

Centre for Speech Technology Research
University of Edinburgh, 80, South Bridge
Edinburgh, EH1 1HN UK

ABSTRACT

Clause-internal filled pauses and preceding peak F_0 values for American and British English speakers were analyzed to determine whether the intonation of filled pauses is relative to, or independent of, prior prosodic context. Higher peaks were found to be systematically associated with higher filled-pause values within speakers, supporting the "relative" hypothesis. In modeling this relationship it was found that a linear model, in which filled-pause F_0 was expressed as an invariant (over speakers) proportion of the distance between peak and baseline, produced results nearly identical to those of a two-parameter model in which the coefficients of peak and baseline were allowed to vary freely. Analyses of additional variables showed the model to be less appropriate for filled pauses after sentence-initial peaks, but unaffected by temporal variables

I. INTRODUCTION

Filled pauses such as "uh" and "um" have been observed to have a low F_0 and level or falling tone [1], and more specifically to have an F_0 lower than that of both accented and unaccented neighboring syllables [2]. These findings have implications for models of human speech perception, automatic speech recognition, and linguistic theory. For example, listeners have difficulty in locating filled pauses when monitoring for sentence content [3]; this may occur because filled pauses are intonationally set off from the message stream. Low F_0 could be utilized by recognition systems as a cue to identifying filled pauses, which have proven difficult to recognize [4]. Linguists may be concerned with how to best represent these predictably low-F_0 units in prosodic descriptions of spontaneous speech.

A question relevant to these issues concerns the nature of the relationship between the low F_0 of filled pauses and the prosody of surrounding material. One possibility is that filled pauses are produced at an absolute, speaker-specific F_0 value regardless of location. A second possibility is that the F_0 of filled pauses varies within speaker, but that the variation is unpredictable. A third possibility is that the F_0 of filled pauses for a particular speaker can be predicted at better than chance given knowledge about the prosodic context.

The current study attempts to address this question by examining filled pauses that occur within a syntactic clause, as opposed to those that initiate a speaker's turn or occur between clauses. Since the question of interest concerned the relationship between the F_0 of filled pauses and prosodic context, the most interesting cases to examine would be those that interrupt a prosodic phrase. Conditioning filled pauses on the basis of prosodic environment, however, poses difficulties in that: (1) prosodic theories are not tailored to the description of material surrounding hesitation phenomena; (2) it is not clear what level of prosodic structure would be appropriate to use as the relevant unit for "interruption"; and (3) conditioning upon prosody is potentially circular in that hesitations may themselves influence the prosody of surrounding material.

The scheme adopted was to study filled pauses that occurred within a syntactic clause. Filled pauses were considered to be "within-clause" if lexical material preceding the filled pause strongly predicted continuation of the utterance after the filled pause. As a measure of prosodic context, the value of the closest preceding F_0 peak was used. While not a perfect nor the only possible measure, the closest peak was chosen because it was easy to identify and relevant as a prosodic unit. As a measure of filled pause F_0, we used the beginning F_0 value of the filled pause.

Within-clause filled pauses from speakers of American and speakers of British English, in two different discourse contexts, were examined to evaluate the three alternative hypotheses. The "absolute" hypothesis predicted that filled pauses would occur at a constant, speaker-dependent F_0 value regardless of the preceding peak F_0. The "random" hypothesis predicted that filled-pause F_0 values from a particular speaker would vary in a manner uncorrelated with preceding peak F_0 values. The "relative" hypothesis predicted some form of systematic relationship between the peak and corresponding filled-pause F_0 values.

II. METHOD

2.1. Subjects

Two quite different sets of data were analyzed. The first was a set of 120 clause-internal filled pauses from digitized utterances from 29 speakers (14 male, 15 female) of American English making air travel plans by speaking to a computer. The multisite database is described in detail in [5]. The majority of examples came from "Wizard-of-Oz" systems, in which a human interpreted and responded to requests and thus "recognition" was perfect; a small number came from interaction with a Spoken Language System [6]. The number of clause-internal filled pauses per speaker used in the analyses ranged from 2 to 13; 82 of the examples came from 12 speakers (6 male, 6 female) having 5 or more examples each.

The second set consisted of 87 filled pauses taken from a corpus of six dialogues recorded digitally at the Department of Linguistics at the University of Edinburgh. Dialogues involved the second author and a colleague or acquaintance (the subject); conversations were natural, spontaneous on various topics, with no set task. The subjects were 3 male and 3 female speakers of British English, without strong regional accents, who were unaware of the purpose of recording the conversations. The number of clause-internal filled pauses per speaker used in the analyses ranged from 6 to 28.

2.2. Filled Pauses

The analyses included only those filled pauses that followed lexical material indicating that the utterance could not end before the filled pause. For example, in:

"The league against um animal furs hasn't caught on"

the preposition before the filled pause is "looking for" an object. In some cases, strict syntactic expectancy was not present, but knowledge of the domain strongly predicted continuation, as in:

"Please show me flights flying uh on Sunday."

Such cases were included in the data set. However, filled pauses in examples in which the only predictor of continuation was a lexical filler such as "well" or "ok," another filled pause, or a conjunction, as in:

"He shoots around a lot but um he beat me the last time."

were considered to be essentially clause-external and were not included in the analyses.

2.3. Apparatus

The digitized waveforms were sampled at 8 or 16 kHz and all waveforms and pitch tracks were examined using the Entropic ESPS/Waves+ software on a Sun 4 workstation.

2.4. Procedure

The American and British data were coded independently by the first and second authors, respectively. For each within-clause filled pause having reliable pitch tracks, the researcher recorded five F_0 values: those of the beginning and ending of the filled pause, of the preceding and following peaks, and of the lowest F_0 in the utterance (measured after final lowering). The smallest of the lowest-F_0 observations for each speaker was used as that speaker's estimated baseline F_0. Peak values were restricted to occur on words within the clause containing the filled pause. In most cases, the peak was marked on a syllable perceived to be accented; in a few cases no accented syllable was available and the highest preceding F_0 value was used. Four measures of duration were recorded, including the duration of the filled pause, that of preceding and following silent hesitation pauses (if any), and that of the time (and also the number of syllables) between the preceding peak and the beginning of the filled pause. Additional facts about the type of token were coded, including the sex of the speaker, whether or not the filled pause preceded a repetition, repair, or fresh start, whether or not the preceding peak was marked on a sentence-initial accent, and whether the filled pause was "um" or "uh."

III. RESULTS AND DISCUSSION

Figures 1-4 show data for a male or female speaker from each of the data sets (American and British). Time-normalized F_0 values are shown for the preceding peak, initial filled pause, final filled pause, and following filled peak of multiple examples of filled pauses for the particular speaker. Each speaker's estimated baseline is also indicated.

3.1. Evidence for the "Relative" Hypothesis

The first thing to note about the plots is that, in general, the drop to the filled pause from the preceding peak scales with the peak values, so that higher peaks tend to have higher following filled pauses. This simple assumption was tested using data from all 35 speakers. The highest and lowest preceding peak F_0 values over all examples from a particular speaker were extracted and the associated filled pause values compared in a Sign test. In 34/35 cases, the higher preceding peak value was associated with a higher filled pause value, $p < .0001$. This highly significant result is consistent with the relative hypothesis and inconsistent with the absolute and random hypotheses.

3.2. Modeling the Relationship

A second observation about Figs. 1-4 is that there seems to be a compressive effect for peaks closer to the baseline, with lower peaks producing less of a drop to the filled pause than higher ones. Exceptions to this trend are the filled pauses following the very highest peak examples in Figs. 1, 2, and 4, which do not drop as far as expected. However, these examples form a special class; they correspond to filled pauses following peaks marked on sentence-initial accented syllables which, as discussed later, turn out to behave differently than other clause-internal filled pauses. In addition, there appears to be a lower bound of F_0; filled pauses do not seem to go below the baseline. These observations suggest that filled-pause F_0 cannot be expressed as a simple subtractive or multiplicative function of peak F_0.

Based on these observations, we proposed a simple linear model, in which filled-pause F_0 is the F_0 value occurring at a fixed proportion of the distance between the peak F_0 and the estimated baseline, or:

$$F_0 \text{ (filled pause)} = r\,(F_0 \text{ peak} - F_0 \text{ baseline}) + F_0 \text{ baseline}$$

This is a single-parameter model, since the coefficients of peak and baseline are both determined by r.

We determined the value of r empirically for each filled pause token from the set of American and British speakers with five or more examples each (18 subjects, 169 filled pauses.) Means for tokens broken down by American/British and male/female are shown in Table 1.

Table 1: Values of r

Subject	Number of Speakers	Number of Tokens	Mean r	Standard Deviation of r
American				
Male	6	39	0.596	0.214
Female	6	43	0.626	0.158
British				
Male	3	55	0.607	0.240
Female	3	32	0.636	0.266

Because results for the American and British data were remarkably similar, data were pooled for all further analyses. Although the value of r appears to be slightly higher for women in both groups, this difference was trivial in light of the magnitude of the standard deviations. That mean r values did not differ across sex (highly correlated with baseline F_0) supports the appropriateness of a linear model. It cannot be determined from these data whether all speakers have the same optimal r, or whether that value differs for different speakers; this question awaits analysis of a larger data set.

A linear regression with the constant term suppressed, performed using the raw data from subjects represented in Table 1, and using the mean r determined over the entire set (0.62), yielded a standard error in prediction of 15.41 Hz. This model

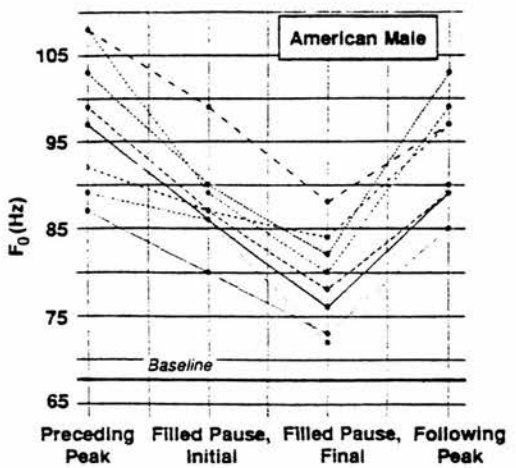

Fig. 1 - Peak and Filled-Pause F_0 for American Male

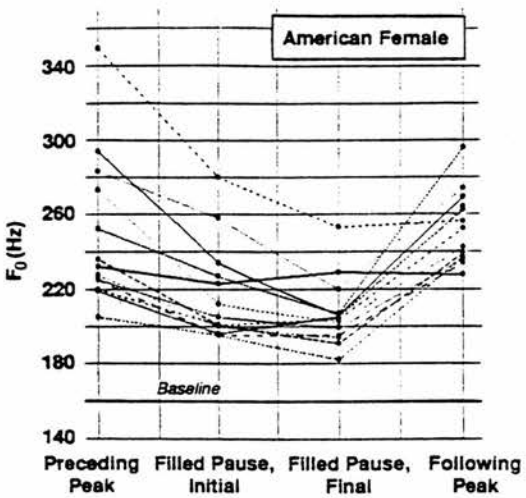

Fig. 2 - Peak and Filled-Pause F_0 for American Female

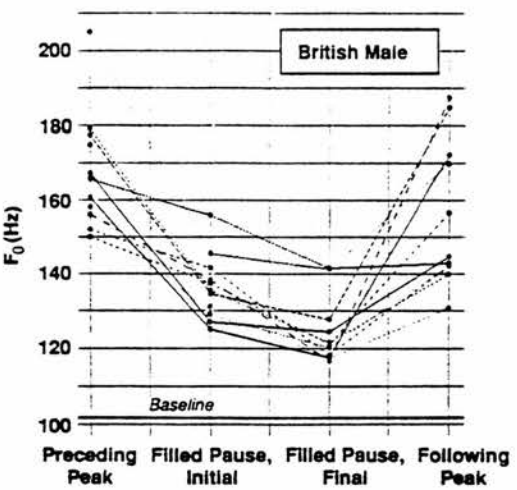

Fig. 3 - Peak and Filled-Pause F_0 for British Male

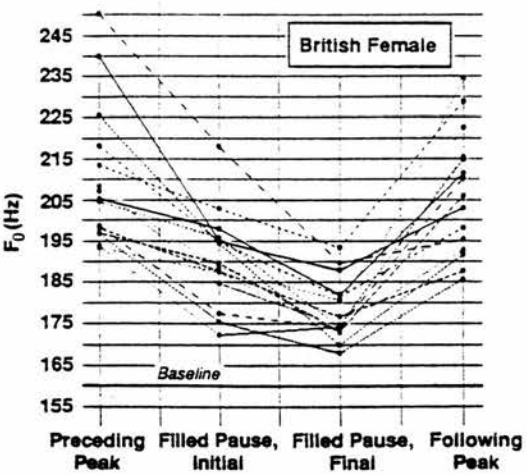

Fig. 4 - Peak and Filled-Pause F_0 for British Female

was clearly better than one in which only the peak was used to predict the filled pause F_0 (standard error = 19.58 Hz). More importantly, the analysis using r and the observed baseline produced a standard error less than 0.2 Hz away from that produced by a two-parameter model (standard error = 15.25 Hz) that allowed the coefficients of peak and baseline to vary freely.

3.3. Optimal Reference F0

An issue addressed was whether, given the single-parameter model, the estimated baseline values used corresponded to the optimal reference F_0 values. Ideally, regressions solving for the optimal r and constant for each speaker would allow for comparison of these results to those obtained using the observed baselines; however, to be meaningful such analyses require more data per speaker. Nevertheless, analyses performed for a subset (N=6) of the 18 subjects who had the largest numbers of examples revealed that in each case the optimal reference F_0 was higher than the observed baseline. Therefore a number of modifications of the observed baseline values in the 18-speaker data set were computed. For each modification, r was redetermined using the new baselines, and filled pauses were predicted using the new, overall average r and new baselines. It was found that the minimum standard error (15.16 Hz, as opposed to 15.41 Hz for the original baselines) was produced when observed baselines were increased by 10%. This result, in which the data were better described using a reference F_0 higher than the measured baselines, is consistent with suggestions that the reference F_0 used to scale pitch over the course of an utterance is higher than the F_0 observed after final lowering [e.g.,7].

3.4. Variables Affecting Prediction Accuracy

Regressions using the observed baseline values and selecting independently for values of additional variables revealed that the factor most influencing prediction accuracy was whether or not the preceding peak was marked on a sentence-initial accented syllable; the standard error for cases involving the sentence-initial accent was 30.3 Hz (N=26) as compared to

10.9 Hz (N=143) for the examples not involving a sentence-initial accent. Cases not involving disfluencies had a lower standard error (14.4 Hz, N=141) than that observed overall (15.41 Hz, N=169); however, results for the different types of disfluencies were inconclusive due to small sample size. Prediction error was not affected by whether the filled pause was "um" or "uh," nor was it affected by the sex of the speaker; that females had a higher standard error (18.42, N=75) than males (12.36, N=94) was expected given the roughly 50% higher baseline values for the females.

Interestingly, there was no correlation between the time or the number of syllables from the peak to the filled pause and the drop size. As shown in Figure 5, the drop in F_0 from the preceding peak to the filled pause did not seem to depend on the amount of time elapsed between these two points. Also noteworthy is the finding that there did not seem to be any relationship between the duration of the filled pause itself and the size of the fall in F_0 over the course of the filled pause, as shown in Figure 6. These results suggest that the intonation of filled pauses may be independent of durational factors.

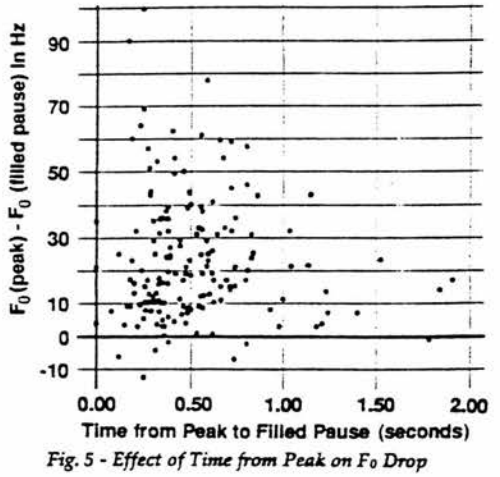

Fig. 5 - Effect of Time from Peak on F_0 Drop

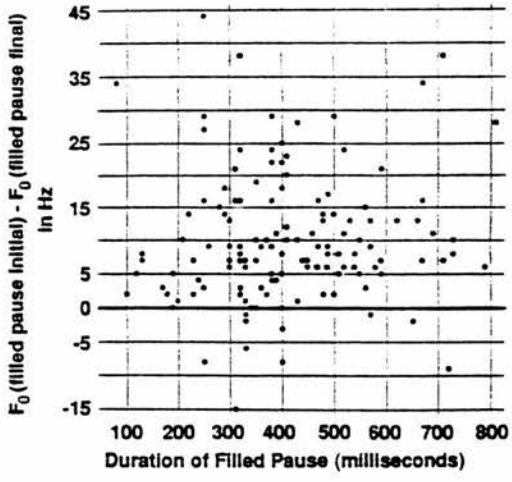

Fig. 6 - Effect of Filled-Pause Duration on Filled Pause Fall

IV. CONCLUSION

The intonation of filled pauses may have implications for models of speech perception, automatic speech recognition, and theoretical descriptions of the prosody of spontaneous speech. We have shown that the F_0 of clause-internal filled pauses does not seem to be absolute or random relative to prior prosodic context, but rather, that the relationship can be described, for both American and British English speakers, by a simple linear model in which the drop is predicted to be an invariant proportion of the distance from the preceding peak F_0 value to the observed baseline. This single-parameter model is remarkably close in prediction accuracy to a model with an additional free parameter. Consistent with theories of the effective reference F_0 during a sentence, the single-parameter model is optimized by reference values higher than those observed sentence-finally. Results also suggest that the intonation of filled pauses may be independent of temporal variables.

ACKNOWLEDGMENTS

The research of the first author was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085 with the Office of Naval Research, and also by a Grant, NSF IRI-890529, from the National Science Foundation.

REFERENCES

[1] O'Shaughnessy, D. D., "Recognition of Hesitations in Spontaneous Speech." *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 521-524, 1992.

[2] Shriberg, E. E., "Intonation of Filled Pauses in Spontaneous Speech." Paper presented at the Conference on Grammatical Foundations of Prosody and Discourse, July 5-6, Santa Cruz, 1991.

[3] Martin, J. G., and W. Strange, "The Perception of Hesitation in Spontaneous Speech." *Perception & Psychophysics*, 3, pp. 427-38, 1968.

[4] Butzberger, J. W., H. Murveit, E. E. Shriberg, and P. J. Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications." *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

[5] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus." *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

[6] Shriberg, E., E. Wade, P. Price, "Human-Machine ProblemSolving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction." *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992

[7] Pierrehumbert, J., "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

PAPER 5

SHRIBERG, E.E., & LICKLEY, R.J. **1992b. The relationship of filled-pause F0 to prosodic context.** *Pages 201–209 of: Proceedings of the IRCS Workshop on Prosody in Natural Speech, Technical Report IRCS-92-37.*

The Relationship of Filled-Pause F0 to Prosodic Context

Elizabeth E. Shriberg

SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025 USA
and Department of Psychology, University of California at Berkeley

Robin J. Lickley

Centre for Speech Technology Research and Department of Linguistics
University of Edinburgh, 80, South Bridge, Edinburgh, EH11HN UK

1. ABSTRACT

Filled pauses in spontaneous speech present problems for models of speech understanding and automatic speech recognition. A potentially important cue to their recognition by both humans and machines is their typically low F0 [9, 7]. The current paper discusses results of a study [10] which sought to determine whether the F0 of filled pauses is relative to, or independent of, the F0 of surrounding lexical material. Clause-internal filled pauses and preceding peak F0 values for speakers of American and British English were examined. Higher peaks were found to be systematically associated with higher filled-pause values within speakers, supporting the "relative" hypothesis. In modeling this relationship it was found that a linear model, in which filled-pause F0 was expressed as an invariant (over speakers) proportion of the distance between the preceding peak F0 and a speaker-dependent terminal low F0, produced results nearly identical to those of a two-parameter model in which the coefficients of peak and terminal low F0 were allowed to vary freely. Analyses of additional variables showed the model to be less appropriate for filled pauses after sentence-initial peaks, but unaffected by temporal variables. These results suggest that clause-internal filled pauses, while lower in F0 than words in the message stream, nevertheless preserve information about the local prosodic context. Implications for psycholinguistics, speech recognition, and linguistic theory are discussed.

2. INTRODUCTION

Phenomena exhibited in spontaneous speech present new challenges for researchers in psychology, speech technology, and linguistics as the object of study shifts from carefully prepared "laboratory speech" to natural conversation. An important difference between spontaneous speech and speech that is read or rehearsed is that spontaneous speech is characterized by relatively high rates of hesitation pauses, repetitions and reformulations [3]. This paper examines one of the most common types of hesitation phenomena: the filled pause, usually realized orthographically as "um" or "uh."

Filled pauses can present problems for models of human language understanding and automatic speech recognition. In the case of human perception, what is remarkable is the extent to which filled pauses are "filtered out" in comprehension. Those familiar with the task of transcribing spontaneous speech will note that filled pauses are often missed in first passes at transcription; laboratory experiments [e.g., 5] have shown that listeners have difficulty locating filled pauses when monitoring for sentence content. In the case of speech recognition, filled pauses are problematic in that they are often misrecognized as words having similar phonetic features, such as "a", "an" or "and," or as syllables of longer words [1, 7, 9].

One source of information that is likely to be important in the successful perception and processing of spontaneous speech in general [see, for example, 6] and speech containing filled pauses in particular, is prosody. Recent work has contributed to our knowledge of the prosodic features of filled pauses. Studies of hesitations in a database of human-computer dialog [4, 11] show that filled pauses tend to occur in the lower region of a speaker's F0 range and have a level or falling tone [7], and, more specifically, that their F0 is typically lower than that of both accented and unaccented neighboring syllables [9].

For human perception, these findings may provide an account for the apparent perceptual separation of filled pauses from the message stream. The low F0 of filled pauses could aid automatic recognizers in distinguishing filled pauses from real words. In addition, linguists may be concerned with how to best represent these predictably low-F0 units in prosodic descriptions of spontaneous speech.

A question relevant to each of these areas concerns the nature of the relationship between the low F0 of filled pauses and the intonation of surrounding material. There are three possible relationships: 1) filled pauses may be produced at an absolute, speaker-specific F0 value regardless of their position within the sentence; 2) the F0 of filled pauses may vary within speaker, but the variation may be unpredictable; or 3) the F0 of filled pauses for a particular speaker may be predictable at better than chance, given knowledge about the prosodic context.

A study previously reported in [10] investigated the relationship between filled-pause F0 and intonational context; the current paper discusses results of that study in further detail. Since the question of interest concerned prosodic context, the relevant filled pauses to examine would be those that interrupt a prosodic phrase, as opposed to those that initiate a speaker's turn or occur between intonation phrases. The task of choosing filled pauses that occur within a prosodic phrase poses difficulties, however, in that: (1) it would be unclear how to label the data prosodically, since existing prosodic theories are not tailored to the description of material surrounding hesitation phenomena; (2) it is not clear what level of prosodic structure would be appropriate to use as the relevant unit for "interruption;" (3) choosing filled pauses on the basis of the prosody of surrounding material is potentially circular in that hesitations may themselves influence the prosody of that material; and (4) prosodic labeling requires listening to utterances and is time-consuming.

The scheme adopted was to study filled pauses that occurred within a syntactic clause. Filled pauses were considered to be "within-clause" if lexical material preceding the filled pause was syntactically incomplete, and strongly predicted continuation of the utterance after the filled pause. The value of the closest F0 peak preceding the filled pause was used as a measure of prosodic context, and the initial F0 value of the filled pause was used as a measure of filled-pause F0.

Within-clause filled pauses from speakers of American and speakers of British English, in two different discourse contexts, were examined to evaluate the three alternative hypotheses. The "absolute" hypothesis predicted that filled pauses would occur at a constant, speaker-dependent F0 value regardless of the value of the preceding peak F0. The "random" hypothesis predicted that filled-pause F0 values from a particular speaker would vary in a manner uncorrelated with preceding peak F0 values. The "relative" hypothesis predicted some form of systematic relationship between the peak and corresponding filled-pause F0 values.

3. METHOD

3.1. Subjects

Two quite different sets of data were analyzed. The first was a set of 120 clause-internal filled pauses from digitized utterances from 29 speakers (14 male, 15 female) of American English making air travel plans by speaking to a computer. The multi-site database is described in detail in [4]. The majority of examples came from "Wizard-of-Oz" systems, in which a human interpreted and responded to requests and thus "recognition" was perfect; a small number came from interaction with a Spoken Language System [11]. The number of clause-internal filled pauses per speaker used in the analyses ranged from 2 to 13; 82 of the examples came from 12 speakers (6 male, 6 female) having 5 or more examples each.

The second set consisted of 87 filled pauses taken from a corpus of six dialogues recorded digitally at the Department of Linguistics at the University of Edinburgh. Dialogues involved the second author and a colleague or acquaintance; they were natural, spontaneous conversations on various topics, with no set task. The subjects were 3 male and 3 female speakers of British English, without strong regional accents, who were unaware of the purpose of recording the conversations. The number of clause-internal filled pauses per speaker used in the analyses ranged from 6 to 28.

3.2. Filled Pauses

The goal of the study was to examine filled pauses that were likely to interrupt a prosodic phrase; however, because it would have been difficult and time-consuming to label the data sets prosodically in order to select the desired filled pauses, a method based largely on syntax was used. In general, the filled pauses selected for analysis were those that directly followed lexical material that would have been syntactically incomplete if the utterance had not continued after the filled pause. It was felt that this would be an efficient, straightforward, and easy-to-replicate method for capturing many of the filled pauses that did interrupt prosodic phrases, while avoiding the complex and time-consuming task of prosodic labeling. Some examples from the American data set are listed in Table 1.

Table 1: Examples of Clause-Internal Filled Pauses

Incomplete	"Looking for"	Example
NP	N	...the lowest [uh] fare...
VP (trans)	NP	...book [uh] the flight...
PP	NP	...leave at [um] noon...
AUX	S	Does [uh] Delta fly...

The researchers tried to determine whether or not a listener would feel it was possible that the speaker could have ended an utterance before the filled pause, based on a transcription alone, but taking semantic and pragmatic information into account. For example, filled pauses in utterances such as:

Show me flights flying [uh] from Boston.

in which material before the filled pause is not necessarily syntactically incomplete, but which would seem incomplete to a listener given the discourse context, were included in the analyses.

Conversely, some utterances which could be viewed as meeting the syntactic expectancy requirement were not included in the analyses. These were cases in which the only item preceding the filled pause in the same clause was a conjunction such as "and" or "but.," a lexical filler such as "well" or "okay," or another filled pause. Such cases were excluded because of the higher likelihood of a prosodic boundary immediately preceding the filled pause.

3.3. Apparatus

The digitized waveforms were sampled at 8 or 16 kHz and all waveforms and pitch tracks were examined using the Entropic ESPS/Waves+ software on a Sun 4 workstation.

3.4. Procedure

The American and British data were coded independently by the first and second authors, respectively. For each within-clause filled pause having reliable pitch tracks, the researcher recorded five F0 values, four measures of duration, and values for four additional variables.

The F0 of each filled pause was measured at both the beginning and end of the filled pause. These values describe the F0 of filled pauses well, since most fall fairly linearly. Analyses in the present work used the initial filled-pause F0 as a measure of filled-pause F0. F0 was also recorded at the F0 peaks most closely preceding and following the filled pause; results reported here used only the preceding peak as a measure of prosodic context. Alternative measures of context (for example topline, or preceding low accents) could also be used, but could be more difficult to measure and locate than F0 peaks. Peak values were restricted to occur on words within the clause containing the filled pause. In most cases, the peak was marked on a syllable perceived to be accented; in a few cases no accented syllable was available and the highest preceding F0 value was used.

A fifth F0 value, which will be referred to as the "terminal low F0," was measured after final lowering in a manner similar to that described in [2]; i.e. for utterances containing a terminal fall, F0 was measured at the lowest point in the fall, disregarding regions associated with errors in pitch tracking or vocal fry. The purpose of this measure was to provide a single, stable, speaker-dependent F0 value for each speaker. The underlying assumption in the present work was that this value should correspond to a speaker's lowest possible F0, as opposed to the lowest F0 realized in any particular utterance, since the former would be the more stable value given the inherently positively skewed distribution of terminal low F0 values. Therefore, terminal low F0 values were obtained for all utterances for a particular speaker that contained a terminal fall. The lowest of these values was then used as the estimate of the speaker's terminal low F0 for all speech tokens from that speaker in the analyses. Care was taken to assure that the lowest terminal F0 value did not appear to be an outlier when compared with the other terminal F0 values obtained for the same speaker.

Four measures of duration were recorded, including the duration of the filled pause, that of preceding and following silent hesitation pauses (if any), and that of the time (and also the number of syllables) between the preceding peak and the beginning of the filled pause.

Values for additional variables of interest were also recorded, including the sex of the speaker, whether or not the filled pause preceded a repetition, repair, or fresh start, whether or not the preceding peak was marked on a sentence-initial accent, and whether the filled pause was "um" or "uh."

4. RESULTS

Figures 1-4 show data for a male or female speaker from each of the data sets (American and British). Time-normalized F0 values are shown for the preceding peak F0, initial filled-pause F0, final filled-pause F0, and following peak F0 in multiple examples of filled pauses for the particular speaker. Each speaker's estimated terminal low F0 is also indicated.

4.1. Testing the Hypotheses: Sign Test

The first thing to note about the plots is that, in general, the drop to the filled pause from the preceding peak scales with the peak values, so that higher peaks tend to have higher following filled pauses. This simple assumption was tested using data from all 35 speakers. The highest and lowest preceding peak F0 values over all examples from a particular speaker were extracted and the associated filled pause values compared in a Sign test. In 34/35 cases, the higher preceding peak value was associated with a higher filled pause value, $p < .0001$. This highly significant result is consistent with the relative hypothesis and inconsistent with the absolute and random hypotheses.

4.2. Modeling the Relationship

A second observation about Figs. 1-4 is that there appears to be a lower bound of F0: filled pauses do not seem to go below the terminal F0. This suggests that filled-pause F0 cannot be expressed as a simple subtractive function of

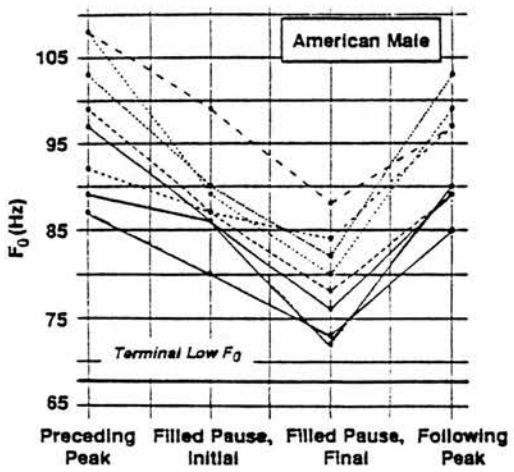

Figure 1: Peak and Filled-Pause F0 for American Male

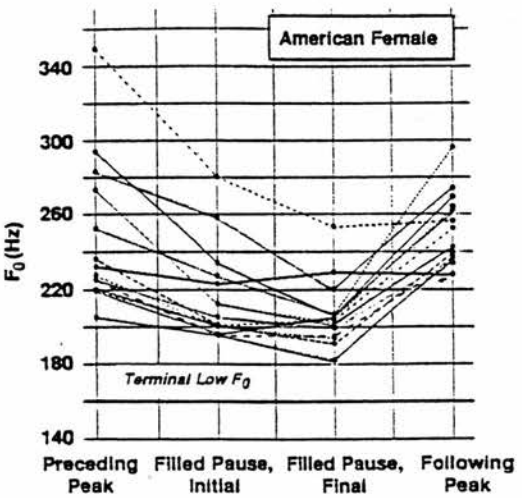

Figure 3: Peak and Filled-Pause F0 for American Female

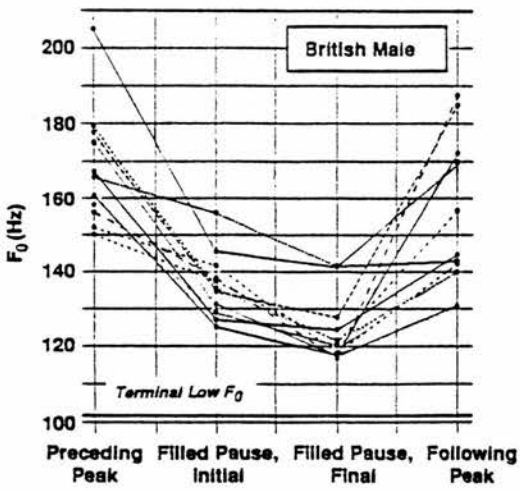

Figure 2: Peak and Filled-Pause F0 for British Male

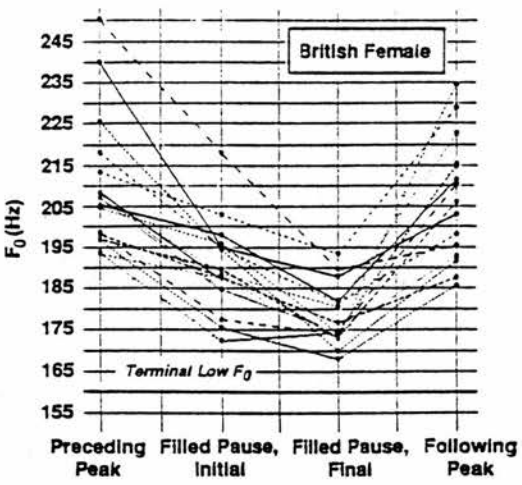

Figure 4: Peak and Filled-Pause F0 for British Female

peak F0. A third observation is that there seems to be a compressive effect for peaks closer to the terminal F0, with lower peaks producing less of a drop to the filled pause than higher ones. This observation suggests that filled-pause F0 cannot be expressed as a simple multiplicative function of peak F0, since such a function would predict parallel curves. Exceptions to this trend are the filled pauses following the very highest peak examples in Figs. 1, 2, and 4, which do not drop as far as expected. However, these examples form a special class; they correspond to filled pauses following peaks marked on sentence-initial accented syllables which, as discussed later, appear to behave differently from other clause-internal filled pauses.

Based on these observations, we proposed a simple linear model, in which filled-pause F0 (F0 fp) is the F0 value occurring at a fixed proportion of the distance between the peak F0 (F0 peak) and the terminal low F0 (F0 min):

$$F_{0\,fp} = r\,(F_{0\,peak} - F_{0\,min}) + F_{0\,min}$$

This is a single-parameter model, since the coefficients of peak F0 and terminal low F0 are both determined by r.

We determined the value of r empirically for each filled pause token from the set of American and British speakers with five or more examples each (18 subjects, 169 filled

pauses.) Means for tokens broken down by American/British and male/female are shown in Table 2.

Table 2: Values of r

Subject	# of speakers	# of tokens	Mean r	s.d. of r
American				
male	6	39	.596	.214
female	6	43	.626	.158
British				
male	3	55	.607	.240
female	3	32	.636	.242

Because results for the American and British data were remarkably similar, data were pooled for all further analyses. Although the value of r appears to be slightly higher for women in both groups, the differences are nonsignificant (as can be seen by comparing them to the magnitude of the standard deviations.)

A linear regression with the constant term suppressed, performed using the raw data from subjects represented in Table 2, and using the mean r determined over the entire set (0.62), yielded a standard error in prediction of 15.41 Hz. A comparison of this model to two other linear models is shown in Table 3. Investigation of higher-order models was not warranted given the lack of evidence for a nonlinear relationship, and the potential danger of over-fitting the small data set at hand. The proposed model was clearly better than one in which only the peak was used to predict the filled pause F0. It was also remarkably close in prediction accuracy to results produced by a two-parameter model which allowed the coefficients of peak and terminal low F0 to vary freely.

Table 3: Comparison of Models

Variables	# of Parameters	RMS error (Hz)
peak, terminal low F0	1	15.41
peak	1	19.58
peak, terminal low F0	2	15.25

4.3. Optimal Reference F0

An issue addressed was whether, given the proposed model, the estimated terminal low F0 values used corresponded to the optimal reference F0 values for prediction. Ideally, regressions solving for the optimal r and constant for each speaker would allow for comparison of these results to those obtained using the observed terminal low values; however, to be meaningful such analyses require more data per speaker. Nevertheless, analyses performed for a subset (N=6) of the 18 subjects who had the largest numbers of examples revealed that in each case the optimal reference F0 was higher than the observed terminal low F0. Therefore a number of modifications of the observed values in the 18-speaker data set were computed. For each modification, r was redetermined using the new terminal low values, and filled pauses were predicted using the new, overall average r and new low F0 values. It was found that the minimum standard error (15.16 Hz, as opposed to 15.41 Hz for the original terminal low values) was produced when observed terminal low values were increased by roughly 10%.

4.4. Effect of Duration

There was no correlation between the time or the number of syllables from the peak to the filled pause and the drop size. As shown in Figure 5, the drop in F0 from the preceding peak to the filled pause did not seem to depend on the amount of time elapsed between these two points.

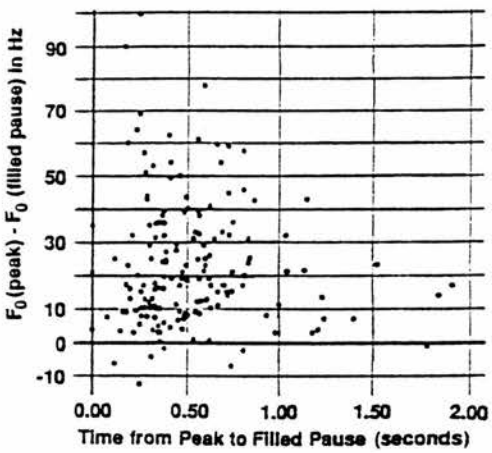

Figure 5: Effect of Time from Peak on F0 Drop

In addition, there did not seem to be any relationship between the duration of the filled pause itself and the size of the fall in F0 over the course of the filled pause, as shown in Figure 6.

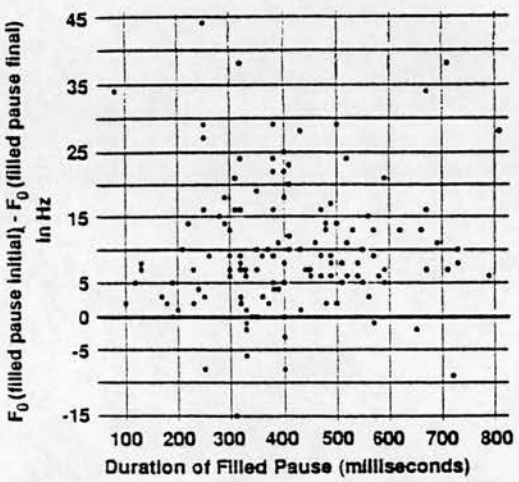

Figure 6: Effect of Filled-Pause Duration on Filled-Pause Fall

4.5. Effect of Additional Variables

Results of regressions performed using the observed terminal low F0 values and selecting independently for values of additional variables are shown in Table 4.

Table 4: Effect of Additional Variables

Data in Analysis	RMS error (Hz)	# of tokens
all data	15.41	169
male speaker	12.36	94
female speaker	18.42	75
peak on sentence-initial accent	30.30	26
peak not on sentence-initial accent	10.90	143
no other disfluency present	14.36	141
filled pause precedes repetition	23.90	11
filled pause precedes replacement	13.09	7
filled pause precedes fresh start	17.90	9
filled pause is "um"	15.29	86
filled pause is "uh"	15.20	83

As can be seen, the factor most influencing prediction accuracy was whether or not the preceding peak was marked on a sentence-initial accented syllable. Although conclusions cannot be drawn given the small number of tokens of this type, it is worth noting that the error in prediction was always in the same direction, with the actual filled pause occurring at a higher F0 value than predicted by the model.- Tokens not involving disfluencies had a lower standard error than that observed overall;, however, results for the different types of disfluencies were inconclusive due to small sample size. Prediction error was not affected by whether the filled pause was "um" or "uh" (although "um" tokens were significantly longer in duration than "uh" tokens, and it should be borne in mind that the present model predicted only the initial F0 of the filled pause.) Prediction accuracy was also not affected by the sex of the speaker; that females had a higher standard error than males was expected given the roughly 50% higher terminal low F0 values for the females.

5. DISCUSSION

5.1. Evaluation of Hypotheses

Two different sets of spontaneous speech data were examined to explore the relationship between the F0 of clause-internal filled pauses and their surrounding context. Results show that the initial F0 of clause-internal filled pauses scales with the F0 of preceding peaks, strongly supporting the "relative" hypothesis.

5.2. Modeling the relationship

Inspection of data from individual subjects revealed that in addition to the scaling of filled pause F0 with preceding peak F0, there was also a lower bound of filled-pause F0 values, and a compressive effect on the size of the drop from the preceding peak to the filled pause as peaks approached the lower portion of a speaker's range.

A model of filled-pause F0 was proposed to reflect these observations. The model was not necessarily intended to have any theoretical interpretation, but rather simply to predict the value of filled-pause F0 using other accessible values of F0. Filled-pause F0 was expressed as a function of three values: (1) a speaker-dependent fixed terminal low F0 value (representing the speaker); (2) the value of the preceding peak F0 (representing the particular prosodic context); and (3) a fixed, speaker-independent scaling factor, r (to express the relationship between the two previous values and filled-pause F0). This is an extremely constrained model, with only one free parameter (r). In addition, the constant term in the model corresponds to a speaker's empirically measured terminal low F0, as opposed to some

F0 value unrelated to prosodic phenomena (for example one outside the speaker's range). Clearly, the current model could also be rewritten to be expressed using coordinates related to a different model (for example, a declination model); the present model is at least as parsimonious as any alternative model in which the functions rewriting peak and terminal low F0 in terms of other variables are linear.

One certainly cannot draw conclusions about the appropriateness of models based on examination of the limited set of data used in the present study. Nevertheless, it is impressive how well the proposed model was able to predict the data. Of possible linear models (there was no evidence for a nonlinear relationship when data from individual subjects were examined) the present model performed extremely well, producing results only very slightly less accurate than a linear model with an additional parameter (in which the coefficients of peak and terminal low F0 were allowed to vary freely.) Real evidence in support of a model such as the present one, however, will probably have to come from comparison of r in the present model to scaling factors proposed in studies of other prosodic phenomena, for example low-tone scaling or the scaling of parentheticals.

5.3. Values of r

It was found that the average value of the parameter r, which expresses the proportion of the distance from terminal low F0 to peak F0 at which filled-pause F0 occurs, did not differ across the American and British data sets. This suggests that the intonation of clause-internal filled pauses, at least as measured by the relationship between preceding peak F0 and initial filled-pause F0, may be independent of factors such as dialect and discourse setting. Mean r values also did not differ across sex. Since speaker sex is highly correlated with the terminal low F0, this lack of a difference in r between sexes is consistent with the appropriateness of a linear model.

5.4. Optimal Reference F0

The value of terminal low F0, a speaker-dependent variable corresponding to the lowest observed F0 value produced after a terminal fall, was found to be slightly lower than the value which optimized prediction. The overall standard error over the data set was slightly decreased when the value of terminal low F0 was raised by 10% for each speaker. A larger data set, with more tokens per speaker, is needed in order to further investigate this finding; it suggests, however, that the value used to scale pitch over the course of an utterance is higher than the F0 measured after final lowering. This is consistent with proposals in the literature [e.g., 8], although it does not distinguish between a declination model and one in which F0 falls abruptly at the end of an utterance. It should be noted that the decision to use the lowest observed terminal low F0, as opposed to other possible values (for example, the mean of all observations) was made because the aim was to get a stable estimate for each speaker, given a positively skewed distribution of low F0 values. Using values such as the mean would therefore be inappropriate. That is, by using mean low F0, one cannot improve results in a principled way, whereas by using a stable estimate such as minimum low F0 (assuming however that there are enough observations available to adequately estimate this value), one can examine the relationship between minimum low F0 and the F0 that optimizes prediction. For exploratory purposes, however, an analysis using mean low F0 values was performed post hoc on the present data set. Results showed a marked reduction in prediction accuracy, and a distribution of r values with much higher standard deviations. Nevertheless, it is conceivable that an analysis using mean low F0 values on a different set of data could produce better results than an analysis using minimum F0 values; such a result would not be meaningful, however, but would rather be due to the fact that mean low F0, like optimal reference F0, is higher than minimum low F0.

5.5. Effect of Duration

Results also suggest that the intonation of filled pauses may be independent of temporal variables. As shown in Fig. 5, there was no correlation between the size of the drop in F0 from the preceding peak to the filled pause and the distance (in time or syllables) between these points; i.e. filled-pause F0 was unrelated to whether or not words and/or silent pauses intervened between the preceding peak and the filled pause. Also, rather surprisingly, there was no correlation between the duration of the filled pause and how far in F0 it fell, as shown in Fig. 6. Most clause-internal filled pauses have a slight linear fall; the fact that longer filled pauses do not fall to a lower F0 than shorter filled pauses implies that the longer tokens either start out with a shallower falling slope, or that they level off in F0 once they reach a point that is "too low" for the local prosodic range. It is also possible that for long hesitations, speakers may stop the filled pause completely and use a silent pause when they have dropped too far. Future work will attempt to examine these issues more closely. These results add further support to the notion that clause-internal filled pauses are in some sense "well-formed" since the range of F0 values for a filled pause is determined by the local prosodic context. In addition, these findings suggest that prosodic regularities in filled pauses may be found more in F0 than in duration measures; this possibility seems reasonable because hesitations, by definition, interrupt the temporal course of production.

5.6. Effect of Sentence-Initial Peaks

As shown in Table 4, prediction error of the proposed model was much greater for filled pauses following peaks marked on sentence-initial accents than for filled pauses elsewhere. In each case following a sentence-initial peak, the prediction of the model for filled-pause F0 was lower than the

observed value; when this relatively small set of tokens was removed from the analyses, the overall error in prediction was reduced substantially. This finding is consistent with the notion that the F0 of filled pauses preserves information about the current prosodic context: filled pauses after peaks corresponding to extra-high sentence-initial accents are themselves extra-high.

5.7. Implications for Areas of Research

The finding that the F0 of filled pauses is relative to prosodic context has implications for models of human speech perception, automatic speech recognition, and for theoretical and descriptive studies of prosody.

The low F0 of filled pauses may help explain why listeners have trouble locating them with respect to words in the message stream; low F0 may also contribute to listeners' ability to filter out filled pauses in comprehension. Experiments designed to test these hypotheses, by using resynthesis to "lift" filled pauses up to the F0 of the region of the lexical material in an utterance, will be conducted in future work. These tests predict that raising the F0 of filled pauses will facilitate listeners' ability to locate them, and also possibly impair comprehension. The finding that the F0 of filled pauses is relative to prosodic context suggests that speakers may attempt to preserve the current prosodic range when hesitating, possibly to inform the listener that they intend to continue where they left off, rather than to abandon a portion of the utterance preceding the filled pause. Thus, a question to be pursued in further work is whether there is a difference between filled pauses that interrupt otherwise fluent clauses, and those that occur at the interruption point of a repair or before a fresh start, since in the latter cases the speaker is abandoning previous material. There were not enough examples of filled pauses in repairs or fresh starts in the present data set to address this question; however preliminary results of additional data suggest that very brief filled pauses, which fall rapidly in F0, often mark a repair (but these are not necessary features for the marking of a repair), and that an unexpectedly high F0 on a filled pause seems to be a very good indicator of a fresh start (essentially an F0 "reset" to begin a new utterance after the filled pause).

Speech recognition systems may be able to take advantage of predictably low F0 in spotting filled pauses. In order to do so successfully however, at least in the case of filled pauses within a clause, these systems will need to take into account the intonation of the local context, rather than using absolute speaker-specific F0 values. Spoken language systems may also benefit from knowing more about prosodic differences between filled pauses in different syntactic environments. Preliminary analyses suggest that whereas clause-internal filled pauses nearly always have a low and falling F0, filled pauses that occur turn-initially or between sentences often have a higher and level or even slightly rising F0. Such information should aid attempts to recognize filled pauses; in addition the recognition of filled pauses having these different prosodic characteristics could contribute information about sentence structure for natural language processing.

As linguists move from the study of read or rehearsed speech to spontaneous discourse, it should become increasingly important for them to consider the prosody of disfluencies, since as shown in the present study, some phenomena considered to be disfluent may exhibit prosodic regularities. This work also suggests that in the case of clause-internal filled pauses, F0, rather than duration, may be the most important prosodic feature to explore. It should prove useful for linguists to include methods for annotating disfluencies in systems developed for the prosodic labeling of spontaneous speech.

6. CONCLUSION

This work has shown that the F0 of one type of speech disfluency, the clause-internal filled pause, is related to the intonation of surrounding material in the message stream. Further work in this area could enhance our knowledge of the production and processing of spontaneous speech, help us learn how to apply these findings to aid speech recognition, and encourage the consideration of hesitations and other disfluencies in theoretical and descriptive work on prosody.

ACKNOWLEDGMENTS

We wish to thank Mark Anderson for helpful discussions on the modeling of F0, and John Bear and Beth Ann Hockey for suggestions regarding syntactic-based principles for categorizing filled pauses. The research of the first author was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085 with the Office of Naval Research, and also by NSF Grant IRI-890529 from the National Science Foundation. The second author was supported by Award number 87310722 from the UK Science and Engineering Research Council. The opinions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

1. Butzberger, J., H. Murveit, E. Shriberg, & P. Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

2. Liberman, M. & J. Pierrehumbert, "Intonational Invariance under Changes in Pitch Range and Length," *Language Sound Structure*, M. Aronoff and R. Oehrle (eds.), MIT Press, 1984.

3. Maclay, H. & C. Osgood, "Hesitation Phenomena in Spontaneous English Speech," *Word*, 15, pp. 19-44, 1959.

4. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

5. Martin, J., and W. Strange, "The Perception of Hesitation in Spontaneous Speech," *Perception & Psychophysics*, 3, pp. 427-38, 1968.

6. Nooteboom, S., P. Brokx & J. De Rooij, "Contributions of Prosody to Speech Perception," *Studies in the Perception of Language*, W. Levelt and F. D'Arcais (eds.), John Wiley and Sons, 1978.

7. O'Shaughnessy, D., "Recognition of Hesitations in Spontaneous Speech." *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 521-524, 1992.

8. Pierrehumbert, J., "The Phonology and Phonetics of English Intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

9. Shriberg, E., "Intonation of Filled Pauses in Spontaneous Speech." Paper presented at the Conference on Grammatical Foundations of Prosody and Discourse, July 5-6, Santa Cruz, 1991.

10. Shriberg, E. & Lickley, R. "Intonation of Clause-Internal Filled Pauses. *Proceedings of the International Conference on Spoken Language Processing*, 1992.

11. Shriberg, E., E. Wade & P. Price, "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992

References

ALLEN, D.E., & GUY, R.F. **1974.** *Conversation analysis: the Sociology of Talk.* **The Hague: Mouton.**

ALLWOOD, J., NIVRE, J., & AHLSÉN, E. **1989. Speech Management: On the Non-Written Life of Speech.** *Gothenburg Papers in Theoretical Linguistics*, **58.**

ALTMANN, G.T.M. **1985.** *Reference and the Resolution of Local Syntactic Ambiguity: The Effect of Context during Human Sentence Processing.* **Ph.D. thesis, University of Edinburgh.**

ANDERSON, A.H., BADER, M., BARD, E.G., BOYLE, E., DOHERTY, G., GARROD, S., ISARD, S., KOWTKO, J., MCALLISTER, J., MILLER, J., SOTILLO, C., THOMPSON, H.S., & WEINERT, R. **1991. The HCRC Map Task Corpus.** *Language and Speech*, **34, 351–366.**

ARAM, D.M., MEYERS, S.C., & EKELMAN, B.L. **1991. On valid and reliable identification of normal disfluencies and stuttering disfluencies - a response.** *Brain and Language*, **40(2), 287–292.**

BARD, E.G., SHILLCOCK, R.C., & ALTMANN, G.T.M. **1988. The Recognition of Words after their Acoustic Offsets in Spontaneous Speech: Effects of Subsequent Context.** *Perception and Psychophysics*, **44(5), 395–408.**

BASHFORD, J.A., REINER, K.R., & WARREN, R.M. **1992. Increasing the intelligibility of speech through multiple phonemic restorations.** *Perception and Psychophysics*, **51(3), 211–217.**

BEACH, C.M. **1991. 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.**

BEAR, J., DOWDING, J., & SHRIBERG, E.E. **1992. 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.**

BLACKMER, E.R., & MITTON, J.L. **1991. Theories of Monitoring and the Timing of Repairs in Spontaneous Speech.** *Cognition*, **39, 173–194.**

BLANKENSHIP, J., & KAY, C. **1964. Hesitation Phenomena in English Speech: a Study in Distribution.** *Word*, **20, 360–372.**

BLESSER, B. **1972. Speech perception under conditions of spectral transformation: Phonetic characteristics.** *Journal of Speech and Hearing Research*, **15, 4–41.**

BUTCHER, A. **1981.** *Aspects of the speech pause: phonetic correlates and communicative functions.* **Ph.D. thesis, Christian-Albrechts-Universität Kiel.**

BUXTON, H. **1983. Temporal Predictability in the Perception of English Speech.** *In: Prosody: Models and Measurements.* **Springer Series in Language and Communication, vol. 14. Berlin: Springer-Verlag.**

CARBONELL, J.G., & HAYES, P.J. **1983. Recovery Strategies for Parsing Extragrammatical Language.** *American Journal of Computational Linguistics*, **9(3-4), 123–146.**

CARDOZO, B.L., & RITSMA, R.J. **1965. Short-term characteristics of periodicity pitch.** *In: Proceedings of the fifth International Congress on Acoustics.* **International Congress on Acoustics, Liège.**

CARLETTA, J., CALEY, R.J., & ISARD, S.I. **1993a (November).** *A Collection of Self-Repairs from the Map Task Corpus.* **Technical Report TR-47. Human Communication Research Centre, University of Edinburgh, Edinburgh.**

CARLETTA, J., CALEY, R.J., & ISARD, S.I. **1993b (March).** *A System Architecture for Simulating Time-Constrained Language Production.* **Technical Report RP-43. Human Communication Research Centre, University of Edinburgh, Edinburgh.**

COLE, R.A., & SCOTT, B. **1973. Perception of temporal order in speech: the role of vowel transitions.** *Canadian Journal of Psychology*, **27, 441–449.**

COOPER, W.E., & PACCIA-COOPER, J.M. **1980.** *Syntax and Speech.* **Cambridge, MA: Harvard University Press.**

COOPER, W.E., SOARES, C., HAM, A., & DAMON, K. **1982. Planning speech for execution at different tempos.** *Journal Of The Acoustical Society Of America*, **72(Suppl. 1), S64.**

COOPER-KUHLEN, E. **1992. Contextualizing discourse: the prosody of interactive repair.** *Pages 337–364 of: The Contextualization of Language.* **Amsterdam: John Benjamins.**

CORDES, A.K., GOW, M.L., & INGHAM, R.J. **1991. On valid and reliable identification of normal disfluencies and stuttering disfluencies - a response.** *Brain and Language*, **40(2), 282–286.**

COTTON, S., & GROSJEAN, F. **1984. The gating paradigm: A comparison of successive and individual presentation formats.** *Perception and Psychophysics*, **35, 41–48.**

CRAIN, S., & STEEDMAN, M. **1985. On Not Being Led up the Garden Path: the Use of Context by the Psychological Parser.** *In: Natural Language Parsing: Psychological, Computational and Theoretical Perspectives.* **Cambridge University Press.**

CUTLER, A. **1976. Phoneme-monitoring reaction time as a function of preceding intonation contour.** *Perception and Psychophysics*, **20(1), 55–60.**

CUTLER, A. **1982.** *Slips of the Tongue and Language Production.* **The Hague: Mouton.**

CUTLER, A. **1983. Speakers' Conceptions of the Function of Prosody.** *In: Prosody: Models and Measurements.* **Berlin: Springer-Verlag.**

CUTLER, A., & BUTTERFIELD, S. **1992. Rhythmic cues to speech segmentation - evidence from juncture misperception.** *Journal of Memory and Language*, **31, 218–236.**

CUTLER, A., & LADD, D.R., (EDS.). **1983.** *Prosody: Models and Measurements.* **Springer Series in Language and Communication, vol. 14. Berlin: Springer-Verlag.**

CUTLER, A., & MEHLER, J. **1993. The periodicity bias.** *Journal of Phonetics*, **21(1-2), 103–108.**

CUTLER, A., & NORRIS, D. **1988. The Role of Strong Syllables in Segmentation for Lexical Access.** *Journal of Experimental Psychology: Human Perception and Performance*, **14(1).**

CUTLER, A., MEHLER, J., NORRIS, D., & SEGUI, J. **1992. The monolingual nature of speech segmentation by bilinguals.** *Cognitive Psychology*, **24(3), 381–410.**

DARWIN, C.J. **1975. On the Dynamic Use of Prosody in Speech Perception.** *Pages 178–193 of:* COHEN, A., & NOOTEBOOM, S.G. **(eds),** *Structure and Process in Speech Perception.* **Berlin: Springer-Verlag.**

DECHERT, H.W., & RAUPACH, M., (EDS.). **1980.** *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* **The Hague: Mouton.**

DEESE, J. **1980. Pauses, prosody, and the demands of production in language.** *In: Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* **The Hague: Mouton.**

DENES, P.B., & PINSON, E.N. **1963.** *The Speech Chain: The Physics and Biology of Spoken Language.* **Baltimore: Bell Telephone Laboratories.**

DICKERSON, W.B. **1971.** *Hesitation Phenomena in the spontaneous speech of non-native speakers of English.* **Ph.D. thesis, University of Illinois at Urbana.**

DORMAN, M.F., CUTTING, J., & RAPHAEL, L.J. **1975. Perception of temporal order in vowel sounds with and without formant transitions.** *Journal of Experimental Psychology: Human Perception and Performance*, **1, 121–129.**

DOWTY, D., KARTUNNEN, L., & ZWICKY, A. (EDS.). **1985.** *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives.* **Cambridge University Press.**

DU BOIS, J.W. **1974. Syntax in Mid-Sentence.** *Berkeley Studies in Syntax and Semantics*, **1, III–1 to III–25.**

DUEZ, D. **1985. Perception of silent pauses in continuous speech.** *Language and Speech*, **28, 377–389.**

DUEZ, D. **1993. Acoustic correlates of subjective pauses.** *Journal of Psycholinguistic Research*, **22(1), 21–39.**

EGIDO, C., & COOPER, W.E. **1980. Blocking of Alveolar flapping in speech production: the role of syntactic boundaries and deletion sites.** *Journal of Phonetics*, **8, 175–184.**

FROMKIN, V.A. (ED.). **1980.** *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* **Academic Press.**

GEE, J.P., & GROSJEAN, F. **1983. Performance Structures: a Psycholinguistic and Linguistic Appraisal.** *Cognitive Psychology*, **15, 411–458.**

GIVON, T., (ED). **1979.** *Discourse and Syntax.* **Syntax and Semantics, vol. 12. New York: Academic Press.**

GOLDMAN-EISLER, F. **1958. Speech production and the predictability of words in context.** *Quarterly Journal of Experimental Psychology*, **10, 96–106.**

GOLDMAN-EISLER, F. **1968.** *Psycholinguistics: Experiments in Spontaneous Speech.* **London: Academic Press.**

GROSJEAN, F. **1980. Spoken word recognition processes and the gating paradigm.** *Perception and Psychophysics*, **28, 267–283.**

GROSJEAN, F. **1983. How Long is the Sentence ? Prediction and Prosody in the On-line Processing of Language.** *Linguistics*, **21(3), 501–529.**

HIEKE, A.E. **1981. A Content-Processing View of Hesitation Phenomena.** *Language and Speech*, **24(2), 147–160.**

HIEKE, A.E., KOWAL, S., & D.C., O'CONNELL. **1983. The trouble with "articulatory" pauses.** *Language and Speech*, **26(3), 203–214.**

HINDLE, D. **1983. Deterministic Parsing of Syntactic Non-Fluencies.** *Pages 123–128 of: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.* **Association for Computational Linguistics.**

HIRSCHBERG, J, & LITMAN, D. **1987. Now Let's Talk about** *Now*: **Identifying Cue Phrases Intonationally.** *Pages 163–171 of: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistic.*

HOCKETT, C.F. **1958.** *A course in modern linguistics.* **New York: Macmillan.**

HOWELL, P., & YOUNG, K. **1991. The Use of Prosody in Highlighting Alterations in Repairs from Unrestricted Speech.** *The Quarterly Journal of Experimental Psychology*, **43A(3).**

INGHAM, R.J., & CORDES. **1992. Interclinic differences in stuttering-event counts.** *Journal of Fluency Disorders*, **17(3), 171–176.**

INGHAM, R.J., CORDES, A.K., & GOW, M.L. **1993. Time-interval measurement of stuttering: Modifying interjudge agreement.** *Journal of Speech and Hearing Research*, **36, 503–515.**

JAMES, D.M. **1972 (14-16 April). Some Aspects of the Syntax and Semantics of Interjections.** *Pages 162–172 of: Papers from the Eighth Regional Meeting of the Chicago Linguistic Society.* **Chicago Linguistic Society.**

JAMES, D.M. **1973 (13-15 April). Another Look at, say, some Grammatical Constraints on, oh ,Interjections and Hesitations.** *Pages 242–251 of: Papers from the Ninth Regional Meeting of the Chicago Linguistic Society.* **Chicago Linguistic Society.**

KASL, S.V., & MAHL, G.F. **1965. The relationship of disturbances and hesitations in spontaneous speech to anxiety.** *Journal of Personality and Social Psychology*, **1, 425–433.**

LABOV, W. **1966. On the Grammaticality of Everyday Speech.** *In: Annual Meeting of the Linguistic Society of America.*

LANGER, H. **1990. Syntactic normalization of spontaneous speech.** *Pages 180–183 of: Proceedings of COLING 90.*

LARKEY, L.S. **1983. Reiterant Speech: an Acoustic and Perceptual Validation.** *Journal Of The Acoustical Society Of America*, **73(4), 1337–1345.**

LASS, R. **1984.** *Phonology: an introduction to basic concepts.* **Cambridge: Cambridge University Press.**

LEHISTE, I., OLIVE, J.P., & STREETER, L.A. **1976. The role of duration in disambiguating syntactically ambiguous sentences.** *Journal Of The Acoustical Society Of America*, **60(5), 1199–1202.**

LEVELT, W.J.M. **1983. Monitoring and Self-Repair in Speech.** *Cognition*, **14, 41–104.**

LEVELT, W.J.M. **1984. Spontaneous self-repairs in speech: Processes and representations.** *Pages 98–101 of:* VAN DEN BROECKE, M.P.R., & COHEN, A. **(eds),** *Proceedings of the 10th ICPhS.* **International Congress of Phonetic Sciences, Dordrecht.**

LEVELT, W.J.M. **1989.** *Speaking: from intention to articulation.* **Cambridge, Massachusetts: The MIT Press.**

LEVELT, W.J.M., & FLORES D'ARCAIS, G.B., (EDS). **1978.** *Studies in the Perception of Language.* **John Wiley and Sons.**

LIBERMAN, M.Y., & STREETER, L.A. **1978. Use of Nonsense-Syllable Mimicry in the Study of Prosodic Phenomena.** *Journal Of The Acoustical Society Of America*, **63, 231–233.**

LICKLEY, R.J., BARD, E.G., & R.C., SHILLCOCK. **1991 (August). Understanding Disfluent Speech: is there an Editing Signal?** *Pages 98–101 of: Proceedings of the ICPhS*, **vol. 4. International Congress of Phonetic Sciences, Aix-en-Provence, France.**

LUZZATI, D. **1987. ALORS: a skimming parser for spontaneous speech processing.** *Computer Speech and Language*, **22, 159–177.**

MACLAY, H., & OSGOOD, C.E. **1959. Hesitation Phenomena in Spontaneous English Speech.** *Word*, **15, 19–44.**

MADCOW. **1992. Multi-site Data Collection for a Spoken Language Corpus.** *In:* MARCUS, M. **(ed),** *Proceedings of the DARPA Speech and Natural Language Workshop.*

MAHL, G. **1957. Disturbances and silences in the patient's speech in psychotherapy.** *Journal of Abnormal and Social Psychology*, **42, 3–32.**

MARCUS, M.P. **1980.** *A Theory of Syntactic Recognition for Natural Language.* **Cambridge, MA: MIT press.**

MARSLEN-WILSON, W.D., TYLER, L.K., WARREN, P., GRENIER, P., & LEE, C.S. **1992. Prosodic effects in minimal attachment.** *Quarterly Journal Of Experimental Psychology Section A – Human Experimental Psychology*, **1, 73–87.**

MARTIN, J.G. **1979. Rhythmic and segmental perception are not independent.** *Journal Of The Acoustical Society Of America*, **65, 1286.**

MARTIN, J.G., & STRANGE, W. **1968. The Perception of Hesitation in Spontaneous Speech.** *Perception and Psychophysics*, **3(6), 427–438.**

MEHLER, J., DOMMERGUES, J.Y., FRAUENFELDER, U., & SEGUI, J. **1981. The Syllable's role in speech segmentation.** *Journal of Verbal Learning and Verbal Behavior*, **20, 298–305.**

MELTZER, R.H., MARTIN, J.G., MILLS, C., IMHOFF, D., & ZOHAR, D. **1976. Reaction time to temporally-displaced phoneme targets in continuous speech.** *Journal of Experimental Psychology: Human Perception and Performance*, **2, 277–290.**

MENS, L.H., & POVEL, D.J. **1986. Evidence Against a Predictive Role for Rhythm in Speech Perception.** *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **38a(2), 177–192.**

MILLER, G.R., & HEGWILL, M.A. **1964. The Effect of Variations in Non-Fluency on Audience Ratings of Source Credibility.** *Quarterly Journal of Speech*, **50, 36–44.**

MOORE, S.E., & PERKINS, W.H. **1990. Validity and reliability of judgments of authentic and simulated stuttering.** *Journal of Speech and Hearing Disorders*, **55(3), 383–391.**

NAKATANI, C., & HIRSCHBERG, J. **1993a. A speech-first model for repair detection and correction.** *In: Proceedings of the ARPA Workshop on Human Language Technology.*

NAKATANI, C., & HIRSCHBERG, J. **1993b. A speech-first model for repair detection and correction.** *Pages 46–53 of: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistic.*

NOOTEBOOM, S.G., BROKX, P.L., & DE ROOIJ, J.J. **1978. Contributions of Prosody to Speech Perception.** *In: Studies in the Perception of Language.* **John Wiley and Sons.**

OLLER, D.K. **1973. The Effect of Position in Utterance of Speech Segment Duration in English.** *Journal Of The Acoustical Society Of America*, **54, 1235–1247.**

O'SHAUGHNESSY, D. **1992a (October). Analysis of false starts in Spontaneous Speech.** *Pages 931–934 of: Proceedings of The ICSLP.*

O'SHAUGHNESSY, D. **1992b. Recognition of Hesitations in Spontaneous Speech.** *Pages 521–524 of: IEEE International Conference on Acoustics, Speech and Signal Processing.*

O'SHAUGHNESSY, D. **1993a. Analysis and automatic recognition of false starts in Spontaneous Speech.** *Pages 724–727 of: IEEE International Conference on Acoustics, Speech and Signal Processing.*

O'SHAUGHNESSY, D. **1993b (September). Locating disfluencies in Spontaneous Speech.** *Pages 2187–2190 of: Proceedings of Eurospeech 93.* **2nd European Conference on Speech Communication and Technology, Berlin, Germany.**

PAUL, S.T., KELLAS, G., MARTIN, M., & CLARK, M.B. **1992. Influence of contextual factors on the activation of ambiguous meanings in context.** *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18(4), 703–717.**

PERKINS, W.H. **1990. What is stuttering?** *Journal of Speech and Hearing Disorders*, **55(3), 307–382.**

PICKETT, J.M., & POLLACK, I. **1963. Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt.** *Language and Speech*, **151–165.**

PIKE, K.L. **1945.** *The intonation of American English.* **Ann Arbor, MI: University of Michigan Press.**

PITT, M.A., & SAMUEL, A.G. **1990. The use of Rhythm in attending to speech.** *Journal of Experimental Psychology: Human Perception and Performance*, **16(3), 564–573.**

POLLACK, I., & PICKETT, J.M. **1963. The intelligibility of excerpts from conversation.** *Language and Speech*, **6, 165–171.**

POLLACK, I., & PICKETT, J.M. **1964. Intelligibility of excerpts from fluent speech: Auditory vs structural context.** *Journal of Verbal Learning and Verbal Behavior*, **3, 79–84.**

POSTMA, A., & KOLK, H. **1992. The effects of noise masking and required accuracy on speech errors, disfluency and self-repairs.** *Journal of Speech and Hearing Research*, **35, 357–544.**

POSTMA, A., & KOLK, H. **1993. The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies.** *Journal of Speech and Hearing Research*, **36, 472–487.**

POSTMA, A., KOLK, H., & POVEL, D-J. **1990. On the relation among speech errors, disfluencies and self-repairs.** *Language and Speech*, **33(1), 19–29.**

POSTMA, A., KOLK, H., & POVEL, D-J. **1991. Disfluencies as resulting from covert self-repairs applied to internal speech errors.** *In: Speech motor control and stuttering.* **Amsterdam: Elsevier Science Publishers B.V.**

PRICE, P.J., OSTENDORF, M., SHATTUCK-HUFNAGEL, S., & FONG, C. **1991. The use of prosody in syntactic disambiguation.** *Journal Of The Acoustical Society Of America*, **90(6), 2956–2970.**

QUENÉ, H. **1992. Durational Cues for word segmentation in Dutch.** *Journal of Phonetics*, **20, 331–350.**

QUINTING, G. **1971.** *Hesitation Phenomena in Adult Aphasic and Normal Speech.* **The Hague, Paris: Mouton.**

REICH, S.S. **1975.** *The function of pauses for the decoding of speech.* **Ph.D. thesis, University of London.**

REPP, B.H. **1992. Perceptual restoration of a missing speech sound: auditory induction or illusion.** *Perception and Psychophysics*, **51(1), 14–32.**

ROCHESTER, S.R. **1973. The Significance of Pauses in Spontaneous Speech.** *Journal of Psycholinguistic Research*, **2, 51–81.**

SAMUEL, A.G., & RESSLER, W.H. **1986. Attention within auditory word perception: Insights from the phonemic restoration illusion.** *Journal of Experimental Psychology: Human Perception and Performance*, **12(1), 70–79.**

SCHEGLOFF, E.A. **1979. The Relevance of Repair to Syntax-for-Conversation.** *In:* GIVON, T. (ed), *Discourse and Syntax.* **Syntax and Semantics, vol. 12. New York: Academic Press.**

SCHEGLOFF, E.A., JEFFERSON, G., & SACKS, H. **1977. The preference for self-correction in the organisation of repair in conversation.** *Language*, **53, 361–382.**

SCHERER, K.R. **1979. Personality Markers in Speech.** *In:* SCHERER, K.R., & GILES, H. (eds), *Social Markers in Speech.* **Cambridge University Press.**

SCHERER, K.R., & GILES, H., (EDS.). **1979.** *Social Markers in Speech.* **Cambridge University Press.**

SCHIFFRIN, D. **1987.** *Discourse markers.* **Cambridge: Cambridge University Press.**

SCOTT, D.R. **1982. Duration as a cue to perception of a phrase boundary.** *Journal Of The Acoustical Society Of America*, **71(4), 996–1007.**

SCOTT, D.R., & CUTLER, A. **1986. Segmental Phonology and the Perception of Syntactic Structure.** *Journal of Verbal Learning and Verbal Behaviour*, **23, 450–466.**

SEBASTIAN, N. **1992. Speech segmentation in Catalan and Spanish - the role of syllables.** *International Journal of Psychology*, **27(3-4), 57.**

SELKIRK, E.O. **1980. The role of prosodic categories in English word stress.** *Linguistic Inquiry*, **11, 563–605.**

SHIELDS, J.L., MCHUGH, A., & MARTIN, J.G. **1974. Reaction time to phoneme targets as a function of rhythmic cues in continuous speech.** *Journal of Experimental Psychology*, **102(2), 250–255.**

SHRIBERG, E.E., & LICKLEY, R.J. **1992a (October 12-16). Intonation of clause-internal filled pauses.** *Pages 991–994 of: Proceedings of the International Conference on Spoken Language Processing*, **vol. 2.**

SHRIBERG, E.E., & LICKLEY, R.J. **1992b. The relationship of filled-pause F0 to prosodic context.** *Pages 201–209 of: Proceedings of the IRCS Workshop on Prosody in Natural Speech, Technical Report IRCS-92-37.*

SHRIBERG, E.E., & LICKLEY, R.J. **1993. Intonation of clause-internal filled pauses.** *Phonetica*, **50, 172–179.**

SHRIBERG, E.E., BEAR, J., & DOWDING, J. **1992. Automatic Detection and Correction of Repairs in Human-Computer Dialog.** *In:* MARCUS, M. **(ed),** *Proceedings of the DARPA Speech and Natural Language Workshop.*

SIEGMAN, A.W. **1979. Cognition and Hesitation in Speech.** *In: Of Speech and time: Temporal Speech Patterns in Interpersonal Contexts.* **New York: Hillsdale.**

SIEGMAN, A.W., & FELDSTEIN, S., (EDS). **1979.** *Of Speech and time: Temporal Speech Patterns in Interpersonal Contexts.* **New York: Hillsdale.**

STREETER, L.A. **1982. Acoustic determinants of phrase boundary perception.** *Journal Of The Acoustical Society Of America*, **64(4).**

SVARTVIK, J., & QUIRK, R. **1980.** *A Corpus of English conversation.* **Lund: C.W.K. Gleerup.**

TAYLOR, T.J., & CAMERON, D. **1987.** *Analysing Conversation: rules and units in the structure of talk.* **Oxford: Pergamon.**

TOUGAS, Y., & BREGMAN, A.S. **1990. Auditory streaming and the continuity illusion.** *Perception and Psychophysics*, **47(2).**

TYLER, L.K., & WARREN, P. **1987. Local and Global Structure in Spoken Language Comprehension.** *Journal of Memory and Language*, **26, 638–657.**

TYLER, L.K., & WESSELS, J. **1983. Quantifying contextual contributions to word-recognition processes.** *Perception and Psychophysics*, **34.**

TYLER, L.K., & WESSELS, J. **1985. Is gating an on-line task? Evidence from naming latency data.** *Perception and Psychophysics*, **38.**

VAN WIJK, C., & KEMPEN, G. **1987. A Dual System for Producing Self Repairs in Spontaneous Speech: Evidence from Experimentally Elicited Corrections.** *Cognitive Psychology*, **19(4), 403–440.**

WARREN, R.M. **1984. Perceptual restoration of obliterated sounds.** *Psychological Bulletin*, **96, 371–383.**

WINGFIELD, A. **1975. The Intonation-Syntax Interaction: Prosodic Features in Perceptual Processing of Sentences.** *Communication and Cybernetics*, **11. In Cohen and Nooteboom (eds).**

WINGFIELD, A., & KLEIN, J.F. **1971. Syntactic structure and acoustic pattern in speech perception.** *Perception and Psychophysics*, **9(1A), 23–25.**

YNGVE, V.H. **1973 (13-15 April). I Forget What I Was Going to Say.** *Pages 688–699 of: Papers from the Ninth Regional Meeting of the Chicago Linguistic Society.* **Chicago Linguistic Society.**