To 'errrr' is human: ecology and acoustics of speech disfluencies

Elizabeth Shriberg
Speech Technology and Research Laboratory,
SRI International, Menlo Park, CA
ees@speech.sri.com

Unlike read or laboratory speech, spontaneous speech contains high rates of disfluencies (e.g. repetitions, repairs, filled pauses, false starts). This paper aims to promote 'disfluency awareness' especially in the field of phonetics – which has much to offer in the way of increasing our understanding of these phenomena. Two broad claims are made, based on analyses of disfluencies in different corpora of spontaneous American English speech. First, an *Ecology Claim* suggests that disfluencies are related to aspects of the speaking environments in which they arise. The claim is supported by evidence from task effects, location analyses, speaker effects and sociolinguistic effects. Second, an *Acoustics Claim* argues that disfluency has consequences for phonetic and prosodic aspects of speech that are not represented in the speech patterns of laboratory speech. Such effects include modifications in segment durations, intonation, voice quality, vowel quality and coarticulation patterns. The ecological and acoustic evidence provide insights about human language production in real-world contexts. Such evidence can also guide methods for the processing of spontaneous speech in automatic speech recognition applications.

1 Introduction

Most research on speech is based on idealized data – read or laboratory speech. Such speech is 'impoverished' with respect to phenomena that occur in the speech we use every day. A clear difference between spontaneous speech and read or laboratory speech is that the former contains significant rates of disfluencies (e.g. filled pauses, repetitions and repairs).

There is good reason to study disfluencies, in both theoretical and applied fields. They are frequent – affecting up to ten percent of words and over one third of utterances in natural conversation. For the study of human language, disfluencies provide a window onto underlying processes affecting human speech and language production. On the applied side, disfluencies present a challenge for automatic speech processing, especially since speech recognition models are often trained on read or highly constrained speech (Butzberger et al. 1992).

The high-level goal of this paper is to increase 'disfluency awareness' for the study

DOI:10.1017/S0025100301001128

of spontaneous speech and language, especially in the field of phonetics – a field that has much to offer in the way of increasing our understanding of these phenomena. Specifically, the paper advances two claims:

- *Ecology Claim*: Disfluencies are related to aspects of the speaking environment associated with speaking spontaneously.
- *Acoustics Claim*: Disfluency affects patterns of observed speech and language at all levels, including acoustics (segmentals and prosody).

These claims, and the evidence in support of them, have implications for the study of human spontaneous speech production, and also for automatic processing of spontaneous speech by machine.

2 Speech data and disfluency types

2.1 Speech data

Evidence in support of the two claims comes from analyses of three corpora of spontaneous American English speech from different domains: (1) human-human free conversation (Switchboard), (2) human-human air travel dialogs (AMEX), and (3) human-computer air travel dialogs (ATIS). An advantage of using spontaneous speech from such corpora is the large amount of transcribed data available, making it possible to study even infrequent disfluency types in their natural environment. A limitation is that one does not have control over the selection of talkers, or of specific factors (social, sociolinguistic, cognitive) that might affect their speech and language production. Nevertheless, given the large amount of data and speakers examined, it is possible to draw some generalizations about disfluencies, which (it is hoped) can promote further, focussed work on specific questions of interest.

The Switchboard corpus contains roughly three million words from over 2430 ten-minute telephone conversations on various topics (Godfrey et al. 1992). Roughly 500 speakers representing all major dialects of American English participated in the task in exchange for a per-call remuneration. Speakers registered by choosing topics of interest (e.g. recycling, sports) from a predetermined set; they were automatically connected to another caller with similar interests. Conversations were therefore between strangers; however, transcribers rated the conversations as highly natural. Various data sets were used for the distributional analyses presented here. The smallest set was based on 40,500 words, 30 speakers, and 2,586 disfluencies. However for the speaker-dependent analyses, considerably more data was used (7,500–40,000 words *per speaker*).

The AMEX corpus of human-human air travel planning dialogs consists of telephone conversations between SRI employees and American Express travel agents (Kowtko & Price 1989). These were 'real' travel planning dialogs; callers had agreed to be recorded, but were calling to make actual travel plans. Only the client's speech was used for the analyses herein because agent speech was highly entrained for this task. The data analyzed for the distributional analyses comprises nearly 13,000 words, 66 different speakers, and 745 disfluencies. The ATIS corpus (MADCOW 1992) is a large, multi-site corpus of human-computer dialog in the air travel planning domain, distributed by the Linguistic Data Consortium. (ATIS stands for 'Air Travel Information System'.) Subjects were given hypothetical travel scenarios and were asked to 'solve' them by speaking to a computer. Distributional analyses in the present work were based on nearly 190,000 words from this corpus, from 523 different speakers, and including 1,694 disfluencies.

Table 1 Disfluency types.

Disfluency type	Example
Filled pause	*uh* – we live in dallas
Repetition	all *the* – *the* tools
Deletion	*it's* – I could get it where I work
Substitution	*any health cover-* – *any health insurance*
Insertion	and *i felt* – *i also felt*
Articulation error	and *[pin]* – *pistachio* nuts

2.2 Disfluency types

Disfluencies were classified as described in Shriberg (1994). The main types are illustrated in table 1, with examples from the Switchboard corpus ('–' indicates the interruption point).

The types are named to reflect the type of change in wording at the lexical level, between the material before the interruption point and that after it. Deletions correspond to what other researchers sometimes call 'false starts'; articulation errors correspond to speech errors (segment exchanges). The typology does not attempt to reflect the cause or function of a disfluency – which turns out to be difficult or impossible to determine – but rather simply to sort disfluencies into types based on observable surface patterns.

3 Ecology

This section addresses the first claim, i.e. that disfluencies are related to factors of the speaking environment in which they occur. Four factors are discussed: the speaking context or 'task', the location in an utterance, the individual speaker, and the speaker and listener gender.

3.1 Task effects

Rates of disfluency per word in spontaneous English speech vary from under 1% for constrained human-computer dialog, to roughly 5–10% for natural conversation (Maclay & Osgood 1959, Levelt 1989, Shriberg 1994, Oviatt 1995, Fox Tree 1995, Clark 1996). For the three corpora examined here, the per-word rate of disfluency was about .06 for both human-human corpora (Switchboard and AMEX), but only about .008 for the human-computer dialog (ATIS). This suggests that conversational partner (human versus computer) is an essential factor in disfluency rate, while whether or not the dialog is goal-oriented (AMEX) or free conversation (Switchboard) appears to have little or no effect on overall rates of disfluency. The suppressed rate of disfluencies in human-computer dialog for ATIS is also partly due to the presence of a push-to-talk mechanism; speakers could first plan their utterances, and when ready could push a button to begin recording. Rates of disfluency in contexts without this option are higher than observed here (Eklund & Shriberg 1998).

Task effects are not only quantitative, but also qualitative: the speaking context also affects the distribution of disfluency TYPES. Figure 1 shows the per-word rate of disfluencies in the three different corpora, broken down by disfluency type.

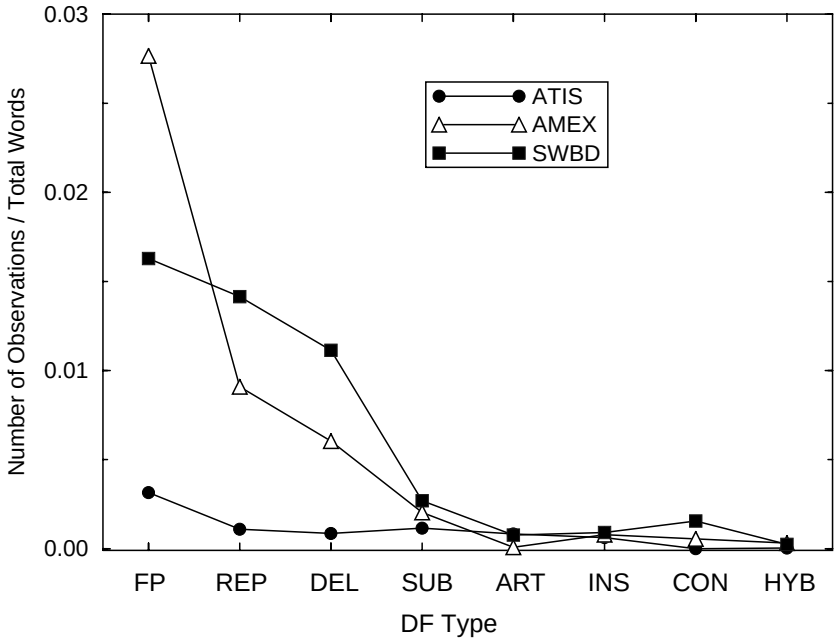

Figure 1 Per-word rate of disfluencies by type. FP = filled pause, REP = repetition, DEL = deletion (also known as a false start), SUB = substitution, ART = articulation error, INS = insertion, CON = repeat or change of coordinating conjunction, HYB = combination of types.

The interesting observation from figure 1 is that the large difference in overall rate of disfluencies between ATIS and the human-human corpora observed earlier is attributable to only three of the disfluency types: filled pauses, repetitions and deletions. The rate of the remaining types is roughly similar for all three corpora.

The rate differences by type suggest that different types of disfluencies arise from different underlying factors. Errors may reflect basic problems in formulation or encoding of a message; the three types that are more prevalent in human-human dialog may function to coordinate exchanges with a conversational partner. Further evidence for this possibility comes from looking at position; the first three types are highly correlated with positions relevant to turn exchanges; i.e. they occur turn or sentence initially. Such results also have implications for modeling disfluencies in automatic speech processing. For human-human dialog, disfluency processing should probably be focused on these three highly frequent types.

3.2 Utterance position effects

The rate of disfluencies is not actually uniform; it depends on utterance position, as shown in figure 2. This result is consistent with past work showing disfluencies to be more likely to occur early in a phrase (Boomer 1965, Beattie 1979) due to increased planning effects. Utterance initial position is also special from the perspective of interaction and turn-taking; work has shown that speakers employ disfluencies such as fillers and repeats to grab or hold the floor, and to secure a listener's gaze before continuing with the message content (Goodwin 1981, Schegloff 1987).

Results also have implications for speech recognition; for example, a language model could assign different prior probabilities to disfluencies, depending on position

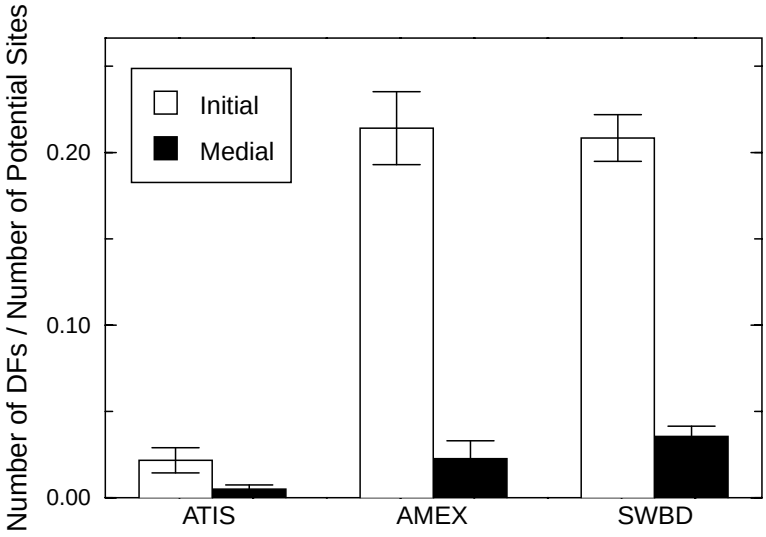

Figure 2 Probability of a disfluency by position in sentence.

relative to a sentence boundary. Interestingly, both initial and medial disfluency rates depend on the length of the sentence; both rates increase for longer sentences. This suggests both increased cognitive processing to produce a longer following sentence, and increased disfluency rates for longer (more complex) sentences.

3.3 Speaker effects

Another important ecological factor related to disfluency production is the individual speaker. Disfluencies rates are shown for 100 different Switchboard speakers in figure 3. The mean of the distribution is about 6 disfluencies per 100 words, as mentioned earlier. However, speakers vary considerably, with some speakers showing about three times the average rate.

Individuals also differ in the relative PROPORTION OF TYPES of disfluencies they produce. Speakers appear to fall into two groups: 'repeaters' produce many more repetitions than deletions, while for 'deleters' the pattern is reversed (see figure 4). The patterns suggest the two types may serve a similar function in conversation.

Furthermore, while speakers may certainly have a stylistic preference for using one pattern or the other, the repeater-deleter difference does not appear to be related *only* to a difference in pattern preference. Rather, it may reflect different heuristics that speakers adopt to cope with the cognitive demands of talking while also planning further speech. The evidence comes from an acoustic measure: speaking rate. Deleters are *faster* speakers than repeaters, in terms of words per unit time. One possible interpretation of this result is that faster speakers 'get ahead of themselves', and thus often have to retract provisional starts and begin anew, whereas slower speakers take more time to plan, increasing hesitations such as repetitions but reducing the need to retract false starts. Results also suggest that it would be useful to represent speakers or speaker-types in automatic disfluency processing, not only to allow for overall rate differences, but also to model differences in type distributions.

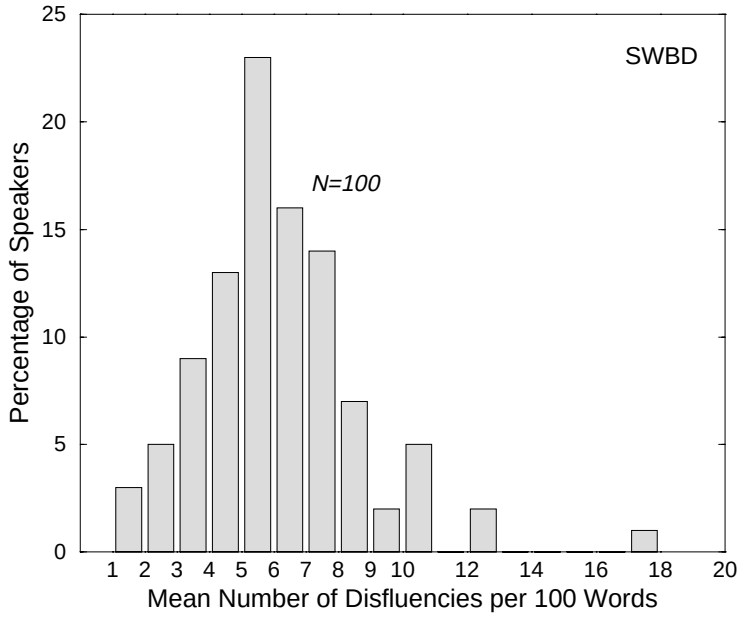

Figure 3 Rate of disfluencies by speaker for 100 switchboard speakers (42 male, 58 female). Disfluency rate is the mean number of disfluencies per 100 effective words for that speaker. Estimates are based on the following average amounts of data PER SPEAKER: 12.37 conversations, 1070 utterances, 7500 effective words, 450 disfluencies.

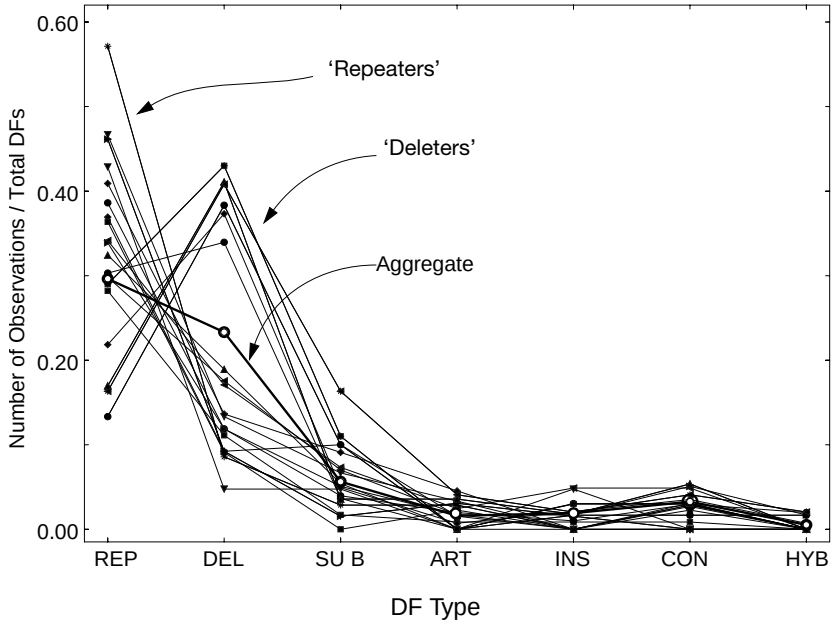

Figure 4 Speaker-specific relative distribution of disfluency types, for 20 Switchboard speakers. 'Repeaters' show a preponderance of repeats; 'Deleters' show a preponderance of deletions. REP = repetition, DEL = deletion, SUB = substitution, ART = articulation error, INS = insertion, CON = coordinating conjunction repetition or substitution, and HYB = combination of two or more of the preceding types.

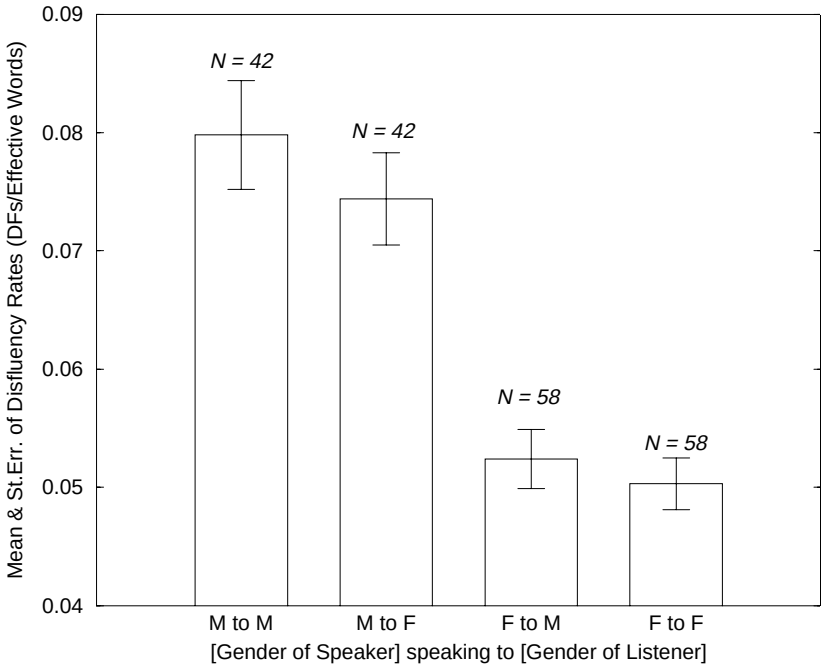

Figure 5 Mean disfluency rate for male and female speakers, by gender of listener.

3.4 Gender and dyad effects

In this final section on ecology, we consider a sociolinguistic factor: gender. Many people seem to assume that women are more disfluent than men. But the Switchboard data show the reverse: men make more disfluencies per word than do women – as shown in figure 5.

It turns out that most of the difference in rates can be attributed to a difference between men and women in the production of filled pauses. An inference is, then, that men may tend to control the floor to a greater extent than women. This does not imply that men spend more time speaking in conversation (since rates are computed per total fluent words), but rather that they may cause their listener to 'wait' for longer than women do. Of course, many further issues remain to be explored here, such as whether or not the effect correlates with social, sociolinguistic or cognitive factors; whether the difference is corpus-dependent, and whether the speaker selection process might have been responsible for the differences. Additional studies, in which social factors are carefully controlled, could shed further light on this question.

An additional interesting observation from figure 5 is that there is also a gender effect associated with the LISTENER – and it is in the same direction as the speaker effect. Speakers (whether male or female) make more disfluencies when talking to men than when talking to women. As with the speaker gender effect, this pattern has a potentially provocative sociolinguistic interpretation. Clearly however, for the less loaded case of automatic speech processing, one could model disfluency rates separately for male and female speakers, and additionally take into account the gender of the listener.

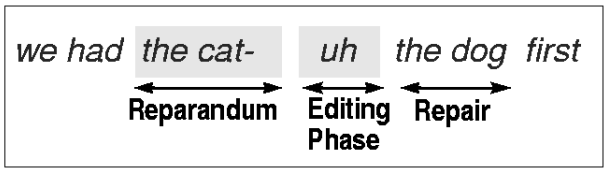

Figure 6 Regions in a disfluency.

4 Acoustics

We turn now to our second claim: disfluency has consequences for the acoustic and phonetic properties of speech. An early suggestion by Hindle (1983) was that disfluencies are marked by a special acoustic 'edit signal' at interruption. Although inspection (Bear et al. 1992), as well as psycholinguistic experiments (Lickley 1994), have revealed no such specific signal, disfluencies are nevertheless associated with a variety of phonetic characteristics that differentiate them from fluent speech. This section first defines the sequential regions by which disfluencies can be commonly described. Subsequently, the discussion walks through each of these regions (from earlier to later in time), describing the types of phonetic modifications that may occur.

4.1 Disfluency regions

The majority of disfluencies that occur in spontaneous speech can be analyzed as having the three-region surface structure presented in figure 6 (Levelt 1983, Bear et al. 1992, Shriberg 1994).

The first region of the disfluency is the REPARANDUM, or material that will later be replaced. The end of this region corresponds to the INTERRUPTION POINT, or the location at which there is a departure from fluency. By this point, the speaker has detected some problem, and according to a 'Main Interruption Rule' halts the production process (Levelt 1983). The editing phase consists of the region from the interruption point to the onset of the repair. This region may be empty, contain a silent pause, or contain editing phrases or filled pauses (*I mean*, *um*, *uh*). The term 'editing' is not intended to imply detection of error; pausing can occur for reasons not involving error. Finally, we have the REPAIR region, which typically reflects the resumption of fluency. (We will assume here that the repair is not itself followed by another self-interruption. If it is, the disfluency is 'complex' (Shriberg 1994).) These regions are contiguous, and removal of the first two (reparandum and editing phase) yields a lexically 'fluent' version.

We can analyze all of our disfluency types this way. A disfluency may contain material only in the editing phase, such as a filled pause. Or it may contain only repeated words in the reparandum and repair. Note that for repeats such as *the the*, this structure predicts that it is the first instance, and not the repeated one, that is most likely to be aberrant, a prediction we will see later evidence for based on phonetic features. Editing terms can combine with different types of disfluency (e.g. *the uh the*; *res- I mean relax*).We will organize our overview of phonetic consequences by moving through these three regions left to right, discussing the effects in each. As we will see, most of the interesting properties occur in the reparandum and editing phases, but certain effects can also be seen in the repair.

4.2 Effects in the reparandum

Although at a lexical level of representation the reparandum is removed in full to arrive at a fluent lexical version, it is not until the speaker notices trouble that we should

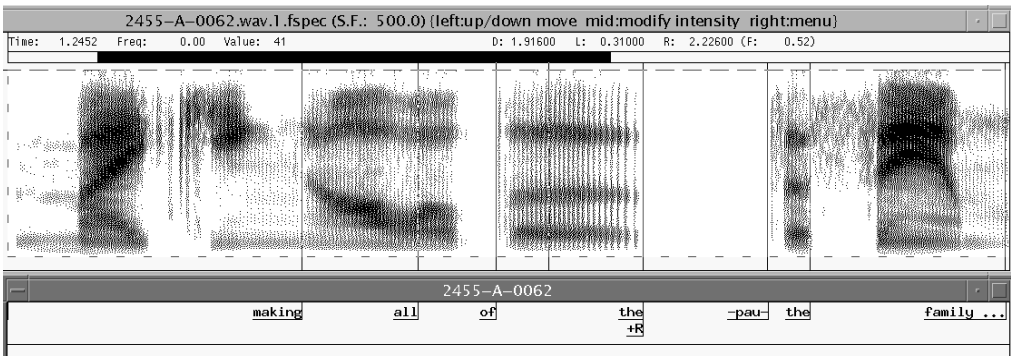

Figure 7 Example of a lengthened reparandum in a repetition of *the the*. Label '-pau-' = pause. Full sentence was *Uh and even those who may not do it seem to be spending more time with their kids and really trying harder at making all of the the family things work.*

expect to see phonetic manifestations. Indeed this is what we find. Phonetic effects in the reparanda of disfluencies are most prevalent at or around the interruption point.

4.2.1 Duration patterns

One of the most commonly observed effects of disfluency is a lengthening of rhymes or syllables preceding the interruption point (Duez 1993, Guaïtella 1993, Lickley 1994, Eklund & Shriberg 1998, Shriberg 1999). For example, in the utterance shown in figure 7, there is a repetition disfluency of *the the* towards the end. As the figure illustrates, the first instance of 'the' (which constitutes the reparandum) is much longer than the second instance.

The second instance, which constitutes the repair, has about the same duration as would be characteristic of fluent production. This suggests that in the reparandum, speakers are signaling delay, hesitating much like they might display with a filled pause.

Lengthening can also be seen earlier than the reparandum, projecting an upcoming repair. As noted earlier, the reparandum is defined via lexical correspondence to the repair. So, for example, in *all of the the family things*, *the* is the reparandum. But if the speaker had uttered *all of the of the family things*, then *of the* would be the reparandum. In either case, we might see lengthening starting for example at *of*, which would be part of the reparandum in the latter case but not in the former. Lengthening also occurs frequently as the only cue to disfluency, i.e. in a lexically fluent sequence (Duez 1993, Guaïtella 1993, Eklund & Shriberg 1998), although such cases were not included in the present study.

The lengthening here is different from the type of pre-boundary lengthening observed in fluent speech. Durationally, the degree of lengthening can be far greater for disfluencies than for fluent boundaries, and the shapes of the distributions are different. The disfluency cases suggest a uniform probability of additional time in a hesitation, while fluent boundaries have a more symmetrical distribution. Second, they are different intonationally, since fluent pre-boundary lengthening is usually associated with a pitch movement conveying a boundary tone (e.g. continuation rise, final fall, etc.), whereas the lengthening accompanying disfluency tends to have a flat or slowly falling pitch contour much like those described in section 4.3.3 for filled pauses. This is not surprising: lengthening, like uttering a filled pause, allows speakers to pause in the production of message content without ceasing phonation.

Of course, not all disfluencies show this behavior, because not all disfluencies are associated with hesitation. For example, as discussed earlier, some disfluencies are

associated with detection of error. In such cases, the reparandum is usually not lengthened, but rather SHORTENED (as discussed below in section 4.2.3).

4.2.2 Voice quality

Durational lengthening in the reparandum is also sometimes accompanied by creaky voice, as shown at the end of the lengthened *the* in figure 7 in the previous section. This gives rise to a 'trailing off' percept, since it is also often accompanied by a decrease in amplitude and drop in pitch. Filled pauses (described in section 4.3), which are often quite long in duration compared to syllables in fluent speech, also have a tendency to end in creaky voice. Interestingly, this modification in voice quality may not operate the same way across languages. In Finnish, for example, it has been proposed that creaky voice may indicate turn-transitional locations (Ogden 2001); in English this does not seem to be the case, since the characteristic often occurs in the reparandum and interruption phase of disfluencies, in which the speaker has not yet completed his or her turn.

4.2.3 Word cutoffs and laryngealization

In read or laboratory speech, we expect words to be completed, but this is not the case in spontaneous speech. Speakers halt production soon after noticing trouble (Levelt 1983), without concern for word boundaries. In the human-computer dialog on air travel planning, nearly 60% of disfluencies contained word cutoffs; rates in the two human-human corpora were about 20–25% (Shriberg 1994). The difference is largely due to the higher relative rate of error repairs in human-computer dialog. Errors are not more frequent overall in such corpora, but because non-error hesitations (filled pauses and repetitions) are suppressed in human-computer dialog with a push-to-talk mechanism for speech input, errors make up a larger proportion of the total disfluencies. Various researchers have described cutoffs as abrupt, often showing some form of laryngealization (Bear et al. 1992, Nakatani & Hirschberg 1994, Lickley 1994, Jasperson 1998). Voice quality modifications have been proposed to serve functions related to managing interaction in conversation (Local & Kelly 1986). Ogden (2001) suggests that glottal stops and other types of supralaryngeal occlusion may be used in Finnish as 'turn-holding'', i.e. 'to project the speaker's claim to the turn'.

Cut off words present a problem for automatic speech recognition because partial-word pronunciations are not usually present in the dictionary. Although one could add all possible initial phone sequences of a word as possible pronunciations, such an approach would create a proliferation of pronunciations that would only degrade performance by increasing confusability. A possible solution is to constrain word fragments to be recognized only as parts of closely following words.

4.2.4 Coarticulation

Another consequence of disfluency is a change in surface coarticulation patterns. In the production of words in fluent speech, articulators generally move toward the articulator positions for the onset of the next word. But in disfluencies, this proximal relationship of coarticulation to actual output word sequence cannot be assumed (Lickley 1994, Shriberg 1999). Coarticulation is governed by the next word in the speaker's phonetic PLAN at the time the word in question is uttered – not by the word sequence that is ultimately produced. In fluent speech, the plan and the final output are consistent, but in disfluencies, following lexical content may be temporarily unavailable, or the plan can change on the fly.

Coarticulation was examined in a study of single-word repeats of *the* and *I* (Shriberg 1999). Note that only the place of articulation can safely be determined for transitions, although there are some cases in which the manner is clear. Table 2 presents results, with examples to indicate transition types.

Table 2 Percentage of coarticulation types in repeats of *the*, with illustrative examples; '–' marks the interruption point.

Transition	Frequency	Example
(a) NONE	722 (88%)	the – the dog
(b) to word after repeat	71 (9%)	the(d) – the dog
(c) to different word	19 (2%)	the(d) – the cat
(d) to repeat itself	3 (.3%)	the(th) – the dog

As shown, most cases of repeats have no detectable final transition. This is different from what is expected in fluent connected speech; here, most cases contained a pause at interruption. For speech recognition models, we may thus want to turn off cross-word modeling at repetition boundaries, or more generally at the interruption point of disfluencies. Case (b), which represents the majority of cases with coarticulation, shows that sometimes disfluency effects can be seen earlier than the location of the element causing trouble. From the transition we can infer that the speaker committed to the word directly after the repetition but stalls earlier, perhaps to keep syntactic or prosodic units intact. Case (c) is almost certainly a covert repair, where some word other than *cat* was caught before it was uttered, and repaired. Case (d) is standard in terms of having a transition consistent with the actual following word, but notice that the following word is the repeat itself. This suggests that in some cases, speakers must be planning to repeat while they are still producing the first instance of the word. As with case (a), cases (b) and (c) also pose problems for cross-word modeling in speech recognition. This time, the problem is that there is acoustic evidence for a segment at the end of the reparandum that is inconsistent with recognizer models constrained to model pronunciation only across contiguous surface words.

An additional result, not shown in the table, is that there appears to be an overall tendency for speakers to produce bilabial closures when hesitating. This can be seen for words that do not end in bilabials (the determination is confounded otherwise), if one looks at rates of observed cases, after adjusting for the rate of cases generously construed as explainable due to a following word with a bilabial onset. We used word trigram sequence probabilities from a statistical language model, and compared observed bilabial transitions to the probability of such transitions given the word history. The language model was trained on about three million words of Switchboard data. Results showed that bilabial gestures were observed significantly more often than predicted by the language model, across different word history contexts. The cause of this behavior is not clear, but perhaps it reflects a tendency to cut off the gesture by closing the lips. It is interesting in this regard that *um* also ends with a bilabial nasal in English and in many other languages with a corresponding filler token. It has been suggested (John Local, personal communication) that the phenomenon might be related to forms like *yep*, *nope* and *welp*, where the bilabial is proposed to serve a closing off function (Heritage & Greatbach 1991, Raymond 2000).

4.2.5 Vowel quality

Disfluency is also associated with alterations in vowel quality. A special case is the word *the*, which has an alternate pronunciation, [ðɪ], before vowel-initial words in many dialects of American English. This variant also shows up in disfluencies (Jefferson 1974, Fox Tree & Clark 1997). An example from Switchboard is shown in figure 8.

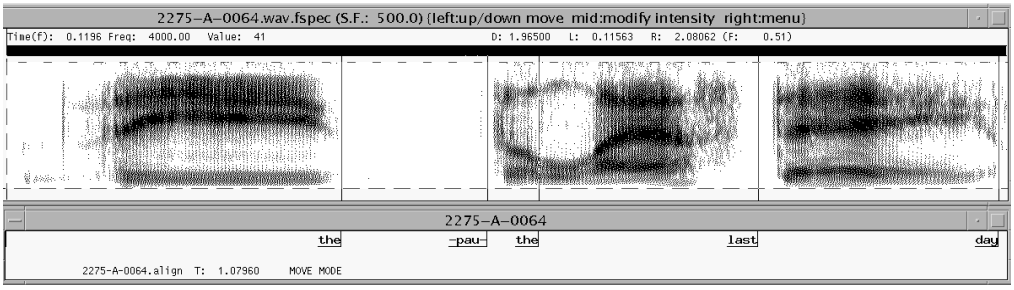

Figure 8 Example of [ði] form in the reparandum of repetition disfluency.

Other words without such variants, but with citation forms that differ from their pronunciation in connected speech, show a similar behavior. For example, *a* and *to* are more likely to be pronounced with their tense vowel forms when in the reparandum of disfluencies, than elsewhere (Shriberg 1999). It is not clear whether such forms are produced as 'signals' to listeners, or whether they reflect a modification related to other acoustic properties such as durational lengthening and following pauses. However, it does seem that speakers choose the alternate form before they start to utter the word, because vowel quality never shifts within the word itself. Such variation is relevant to pronunciation modeling in automatic speech recognition. Many systems have both tense and the lax vowel versions for the pronunciation of these words, but since the lax forms have a higher overall probability of occurrence in spontaneous speech, the tense forms often cause misrecognitions when they occur. For example, in the SRI ATIS system it was not uncommon to observe errors such as *two* for *to* and *eight* for *a* when the tense forms of such words showed up in disfluencies.

4.3 Effects in the editing phase

4.3.1 Unfilled pauses

Under Levelt's framework of speech production (Levelt 1989), self-interruption is associated with a halting of the speech production process at all levels. Therefore, some minimum time is needed after the speech is cut off in order to plan the repair. Disfluency is thus often indicated by unfilled pauses in the editing phase (Goldman-Eisler 1968, Deese 1980, Butcher 1981, Levelt 1989). For automatic speech processing of disfluencies, such pauses have proven to be helpful cues for disfluency detection and correction (Bear et al. 1992, Shriberg et al. 1997, Stolcke et al. 1998, Heeman & Allen 1999). Not all disfluencies, however, contain pauses. For example, Blackmer & Mitton (1991) found that over 30% of the disfluencies in speech to a Canadian radio call-in show had editing phase pause durations below 100 milliseconds, which is less than half the time typically viewed as necessary for a planning pause (Goldman-Eisler 1968). Furthermore, a number of disfluencies showed no pause at all. This suggests that in some cases the plan to repair may occur earlier than the surface interruption point.

4.3.2 Filled pause duration

In English, the vowel in the filled pauses *um* and *uh* is typically close to schwa. However, it can also carry stress, or occur at other regions in the vowel space. In automatic speech recognition, filled pauses are sometimes misrecognized as *a* or as parts of other words containing similar vowels. But filled pauses differ dramatically from these other instances in duration. To illustrate, durations for the vocalic portion

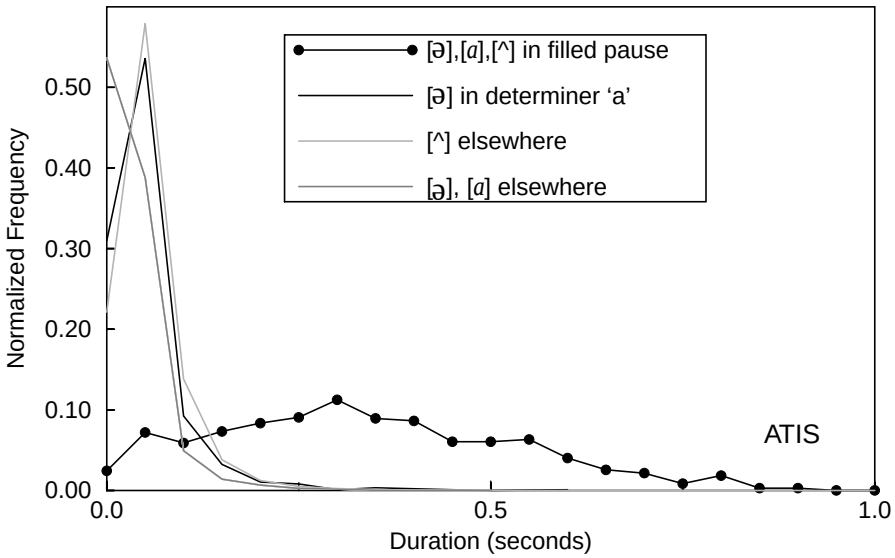

Figure 9 Duration of vowels in filled pauses and elsewhere.

of 700 filled pauses and for 40,000 instances of the same vowels elsewhere, including in the determiner *a*, were obtained from recognizer forced alignments using the ATIS corpus. Results are shown in figure 9.

As shown, vowels in filled pauses have much longer durations than the same vowels in fluent contexts. Duration, then, is a simple cue that could be used by speech recognition systems in discriminating vowels in filled pauses from the same vowels elsewhere. It is also important to treat such durations separately in duration modeling for other purposes, so as not to skew the distributions for the non-filled pause vowels.

4.3.3 Filled pause intonation

Filled pauses have been shown to be low in fundamental frequency (F0),and to display a gradual, roughly linear F0 fall (O'Shaughnessy 1992). In addition, the F0 of filled pauses occurring within a clause was found to be related to the F0 of the surrounding speech (Shriberg & Lickley 1993). Figure 10 shows F0 values for the onset and offset of a filled pause, and the preceding and following F0 peaks. Lines connect points for a specific filled pause. The four F0 measurements are plotted at equally spaced intervals; therefore the actual temporal intervals between these points (which varied greatly) are not represented in the figure. The solid heavy line indicates the speaker's estimated 'baseline' F0, as estimated by measuring F0 at the end of sentence-final F0 falls.

What is striking here is that the F0 of filled pauses falls about halfway between the preceding peak value and the speaker baseline. In fact, F0 values in the study were well predicted by a simple additive-multiplicative model based on these values. These relationships held despite considerable differences in time intervals between the four measured values plotted at regular intervals as in figure 10. These findings suggest speakers may preserve intonational relationships under changes in duration necessitated by the need to pause.

4.3.4 Voice quality: diplophonia

Disfluency is also correlated with diplophonia, a relatively rare form of phonation. This voice quality is produced with a pattern of voice vibration characterized by an

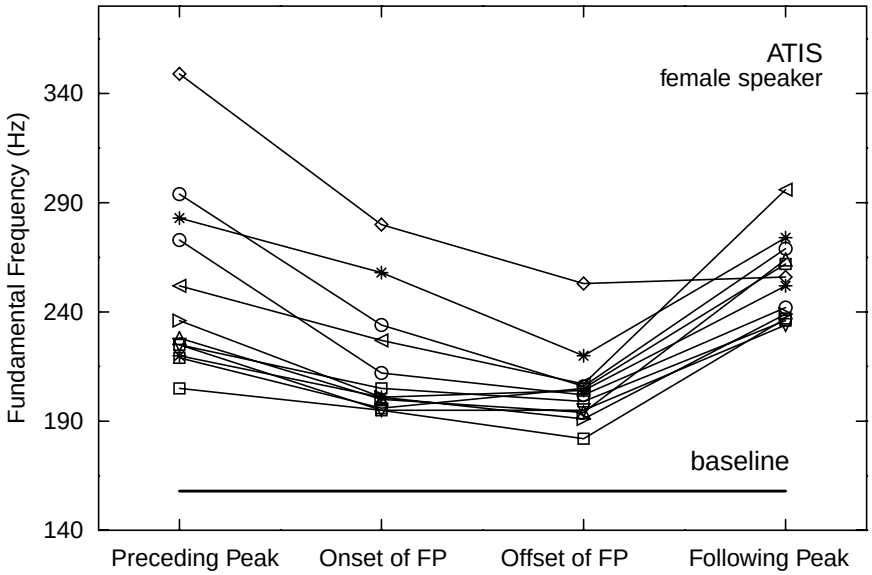

Figure 10 F0 of filled pauses and surrounding peaks.

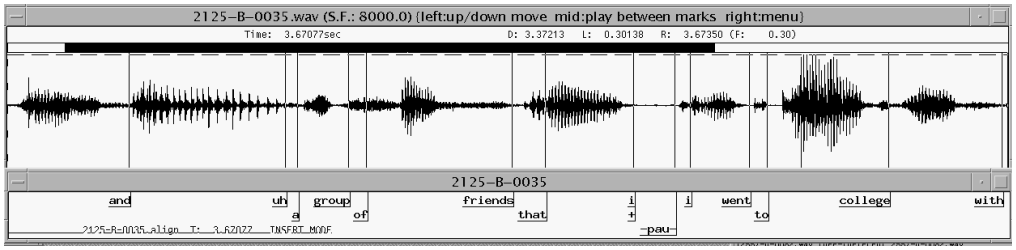

Figure 11 Example of diplophonic phonation (on *uh*).

alternation between strong and weak glottal excitations during phonation. In a study of voice quality we have found that diplophonia occurs exclusively in regions of hesitant speech with drawn out durations, particularly on filled pauses. An example is shown in figure 11.

Diplophonia is often seen on filled pauses, so it is listed here as an effect in the editing phase, but like increased duration and a flat or falling intonation, diplophonia can also be observed at the edge of a reparandum. The example in figure 11 provides another disfluency, a repeat, later in the utterance. Note that while we see reparandum lengthening on the first *I*, there is no evidence of diplophonia in this case.

4.4 Effects in the repair

As said earlier, most consequences of disfluency are located in the reparandum and editing phase, since the repair region constitutes the onset of fluency. An exception, however, is that certain types of repair can show effects of a change in content, in the form of contrastive emphasis on the repairing element (Levelt & Cutler 1983; Howell & Young 1991). Levelt & Cutler (1983) looked at prosodic marking, or an increase in F0,

duration, or amplitude, in the repair region of disfluencies from a pattern description task. They found that marking occurred for roughly half of the repairs involving error, and for only about 20% of the repairs involving mere elaboration. This suggests that it may be more important to call attention to outright error than to inappropriateness. Such marking also illustrates that we cannot simply remove the reparandum and editing phase, leaving a perfectly fluent repair. All three regions are still in the discourse record; the prosodic contrast in the repair is produced with respect to the earlier mention in the reparandum.

5 Conclusion

This paper has aimed to provide a high-level overview of distributional and acoustic properties of speech disfluencies in American English, in support of two claims. The ecology claim – that disfluencies are related to factors associated with the speaking context – was supported by distributional evidence from task effects, disfluency location, speaker effects, and gender effects. The acoustics claim – that disfluency affects phonetic and prosodic patterns of speech – was supported by examples of modifications in duration, voice quality, coarticulation, vowel quality, and intonation patterns.

It is hoped that the paper has served to increase awareness of disfluencies in the study of phonetics. While the detailed distributional and phonetic characteristics of disfluencies are likely to depend on the language and culture studied, it would be interesting to explore whether the two broad claims hold for different varieties of English and in different languages.

Finally, given the added complexity of spontaneous speech over the type of speech observed in the laboratory, the 'holy grail' in linguistics research should include understanding how we communicate in everyday life. Clearly, disfluencies pose a challenge for models of both human and machine processing; a good model in either case must, in the very least, be able to account for the observed data. The many acoustic consequences of speech disfluency suggest that phoneticians could be of great help to psycholinguists and engineers in modeling disfluencies, and more importantly, in modeling spontaneous speech in general.

Acknowledgements

The author gratefully acknowledges Madelaine Plauché for acoustic-phonetic data analysis, and Adrian Simpson, Klaus Kohler, Richard Ogden and John Local for helpful comments and discussion. This research was supported by the National Science Foundation under STIMULATE grant IRI-9619921 and by NASA under NCC2-1256; the views herein are those of the author and should not be interpreted as representing the policies of the funding agency.

References

Bear, J., Dowding, J. & Shriberg, E. E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 56–63.

Beattie, G. W. (1979). Planning units in spontaneous speech: some evidence from hesitation in speech and speaker gaze direction in conversation. *Linguistics* **17**, 61–78.

Blackmer, E. R. & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition* **39**, 173–194.

Boomer, D. S. (1965). Hesitation and grammatical encoding. In Oldfield, R. C. & Marshall, J. C. (eds.), *Language and Speech*, 148–158. Penguin.

Butcher, A. (1981). *Aspects of the Speech Pause: Phonetic Correlates and Communicative Functions*. Ph.D. dissertation, University of Kiel.

Butzberger, J., Murveit, H., Shriberg, E. & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*, 339–343. New York: Morgan Kaufmann.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Deese, J. (1980). Pauses, prosody, and the demands of production in language. In Dechert, H. W. & Raupach, M. (eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, 69–84. The Hague: Mouton.

Duez, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psychological Research* **22(1)**, 21–39.

Eklund, R. & Shriberg, E. (1998). Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human-human and human-machine dialogues. *Proceedings of the International Conference on Spoken Language Processing*, vol. 6, 2631–2634. Sydney: Australian Speech Science and Technology Association.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* **34**, 709–738.

Fox Tree, J. E. & Clark, H. H. (1997). Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* **62**, 151–167.

Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 517–520. San Francisco.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: Academic Press.

Guaïtella, I. (1993). Functional, acoustical and perceptual analysis of vocal hesitations in spontaneous speech. *Proceedings of the ESCA Workshop on Prosody. Working Papers 41*, 128–131. Lund, Sweden: Department of Linguistics and Phonetics.

Heeman, P. & Allen, J. (1999). Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. *Computational Linguistics* **25**, 527–571.

Heritage, J. & Greatbach, D. (1991). On the institutional character of institutional talk: the case of new interviews. In Boden, D. & Zimmerman, D. H. (eds.), *Talk and Social Structure*, 93–137. Berkeley: University of California Press.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, 123–128.

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).

Jasperson, R. (1998). *Repair after Cut-off: Explorations in the Grammar of Focused Repair of the Turn-constructional Unit-so-far*. Ph.D. dissertation, University of Colorado at Boulder, Boulder, CO.

Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society* **2**, 181–199.

Kowtko, J. & Price, P. (1989). Data collection and analysis in the air travel planning domain. *Proceedings of the DARPA Speech and Natural Language Workshop*, *Cape Cod*, 119–125.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition* **14**, 41–104.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*.Cambridge, MA: MIT Press.

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

Lickley, R. J. (1994). *Detecting Disfluency in Spontaneous Speech*. Ph.D. dissertation, University of Edinburgh.

Local, J. & Kelly, J. (1986). Projection and 'silences': notes on phonetic and conversational structure. *Human Studies* **9**, 185–204.

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. *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, 7–14, New York: Morgan Kaufmann.

Nakatani, C. H. & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* **95(3)**, 1603–1616.

Ogden, R. (2001). Turn-holding, turn-yielding and laryngeal activity in Finnish talk-in-interaction. This volume.

O´Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. *ICASSP-92*, vol. 1, 521–524. San Francisco: IEEE.

Oviatt, S. L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* **9**, 19–35.

Raymond, G. (2000). *The Structure of Responding: Conforming and Nonconforming Responses to Yes/No Type Interrogatives*. Ph.D. dissertation, University of California, Los Angeles.

Schegloff, E. A. (1987). Recycled turn beginnings: (a) precise repair mechanism in conversation's turn-taking organisation. In Button, G. & Lee, J. R. E. (eds.), *Talk and Social Organisation*. Clevedon: Multilingual Matters.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California, Berkeley.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences*, vol. 1, 619–622. San Francisco.

Shriberg, E. E. & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica* **50**, 172–179.

Shriberg, E., Bates, R. & Stolcke, A. (1997). A prosody-only decision-tree model for disfluency detection. *Proceedings of the 5th European Conference on Speech Communication and Technology*, vol. 5, 2383–2386.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauché, M., Tür, G. & Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. *Proceedings of the International Conference on Spoken Language Processing*, vol. 5, 2247–2250.