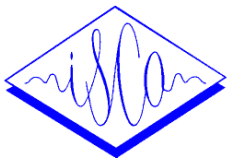

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 21–25.

Prosodic cues of spontaneous speech in French

Katarina Bartkova

France Telecom R&D, Lannion, France

Abstract

Disfluencies, when present in speech signal, can make syntactic parsing difficult. This difficulty is increased when machines are involved in communication and when speech devices rely on automatic speech recognition techniques. In order to improve automatic speech parsing and thus speech comprehension, methods have been proposed to filter disfluencies out from the speech signal. Attempts have been made to use prosodic parameters to improve such a filtering. However, before introducing prosodic parameters into automatic speech recognition processes, it would be useful to investigate whether disfluencies can be characterized in a prosodic way and whether their prosodic cues would be representative enough to be used in automatic systems. The aim of this study was to examine to which extent prosodic parameters would be able to characterize disfluencies in French. Word repetitions, filled and silent pauses and speech repairs were described in a prosodic way using statistical analyses of their prosodic parameters. These analyses allowed simple prosodic rules to be formulated. The efficiency of the prosodic rules was evaluated on the task of filled pauses, word repetitions and hesitation detections.

1. Introduction

Disfluencies, frequently present in spontaneous speech, cause ill-formed and longer sentences which are thus harder to process for natural language understanding. A listener must break the continuous speech stream down into component parts, then build a syntactic structure and determine the meaning that the speaker intended to convey. According to Bailey & al. [3] non-word disfluencies can affect the syntactic parsing and make the syntactic reanalysis more difficult. They also suggest that utterances with disfluencies are parsed differently from those without disfluencies. Thus, although disfluencies have often been viewed as pragmatic phenomena, they can affect language comprehension. According to Bortfeld [6] disfluency is strongly task related and only slightly age related: older speakers produce only slightly higher disfluency rates than young or middle-aged speakers. The task relation of disfluency was confirmed by Arnold et al. [2] who claim that disfluencies occur more often during references to things that are discourse-new, rather than given. Asp [1] found a higher amount of disfluencies in "open" speech acts.

In speech recognition, researchers have investigated techniques to minimise the impact of non-speech events on the performance of the speech recognition systems. R.C. Rose [9] used careful manual labelling of some disfluencies and background events as an additional level of supervision for the training of the acoustic models and the statistical language models. Honal et al. [7] experimented a noisy-channel model which automatically corrected disfluencies using manually transcribed spontaneously spoken speech. Hutchinson et al. [10] presented a similar approach by removing fillers and reparanda and thus transforming utterances into fluent ones. Baron [4] and Shriberg [11] used prosodic parameters to locate disfluencies in spontaneous speech in English. In both studies, prosodic parameters proved to befor disfluency detection.

The aim of the present study is to describe speech disfluencies in French according to their location and prosodic features. This study tries to answer the question whether the prosodic parameters of disfluencies are typical enough to characterise them in a consistent manner. The disfluencies studied here were silent and filled pauses, word repetitions and speech repairs. In order to detect regularities in their prosodic parameters, the values of selected phone durations and the F0 slopes were statistically analysed.

2. Overview

2.1. Speech Data Base Used

The speech data base which was used here consisted of more than 1080 telephone messages in French, left by clients during a survey dedicated to the analysis of clients' satisfaction for phone services. The data base contained 55180 words, i.e. on average each message contained 54 words. The data base was manually transcribed including the annotation of non-speech noises such as inspirations, laughter, background noises… as well as interrupted words, interrupted sentences and filled pauses. This orthographic transcription was used to automatically align the words and their phonetic transcriptions with the speech signal. Thus, prosodic events became accessible such as silent pause occurrences, silent and filled pause durations, F0 values and vowel durations as juncture cues.

2.2. Investigated prosodic parameters

In order to characterize the disfluencies at the prosodic level, F0 patterns and vowel durations were investigated. The vowel durations and the F0 values were measured on the last vowel of the part of the speech signal under consideration. The last vowel from which prosodic parameters were extracted was different from the final schwa like vowel, which can occur in French after each uttered consonant event when not present in the spelling form. The parameters of the schwa vowel were used exclusively when it was the only vowel in the word.

An F0 slope was calculated for each vowel under consideration as the difference between the F0 measured at the end and at the beginning of the vowel, normalized according to the vowel length. In order to facilitate the comparison of the F0 slopes, they were grouped into 5 categories according to their value: flat slope (ranging from -0.3 to +0.3), mid-high slope (ranges from +0.3 to +1.5), high-high slope (higher than 1.5), mid-low slope (between -0.3 and -1.5) and finally low-low slope (lower than -1.5).

To facilitate the vowel length statistical analysis, histograms were calculated grouping the vowel durations into 50 ms intervals.

Silent pauses were categorized according to their duration into three categories: short pauses (shorter or equal to 150 ms), long pauses (longer than 250 ms) and mid-long pauses (situated between short and long pauses).

The measures of the prosodic values (phone durations and F0 slopes) were carried out on the last syllable of the word as the major prosodic cues of the syntactic structuring in French are

located on the last syllable of the word. Martin [9] showed that low and high F0 slopes alternate on the syntactic junctures in the reading style. As far as the vowel duration is concerned, according to Bartkova [5] they are lengthened on the syntactic boundaries as a function of the boundary depth and the right consonantal context. However, the length of the last vowel, located on the syntactic junctures, can remain relatively independent from the other phone durations of the same prosodic group.

3. Disfluency analysis

As mentioned above, the speech disfluencies analyzed in this study were word repetitions, filled and silent pauses and word repairs.

3.1. Word repetition

Word repetitions are frequently encountered in spontaneous speech. Although word repetition was not explicitly labelled in the data base, it was easily retrieved automatically. A word repetition is most of the time the repetition of only **one word 73 %** (688 cases), a sequence of **2 words 23 %** (156 cases) and a sequence of **3 words 4 %** (25 cases) (3 words repetitions are the longest ones found in our data). A same word can be repeated several times in a row (up to 5 times in the corpus studied here). Some word repetitions (especially repetitions of adverbs) can be used for stylistic purposes, to highlight the meaning of the repeated word. The adverb repetitions were kept in our statistics only when separated by long pauses for in such cases the successive adverbs could be considered as being interrupted by a hesitation and were perceived more as a repetition than a stylistic use. Speakers hardly ever repeat lexical words. Repetition mostly concerns grammatical words such as prepositions, articles, auxiliary verbs, pronouns, conjunctions...

One word repetitions

79% of one word repetitions were repetitions of a grammatical word. The most frequently repeated word (18% of all one word repetitions) was the article "**de**". 19% of word repetitions contained adverbs and only 3% contained lexical words.

Two word repetitions

Like in the case of one word repetitions, structures containing exclusively grammatical words such as articles, pronouns, auxiliary verbs, prepositions and conjunctions, constituted the majority, that is 65% of cases. 17% of repeated structures contained adverbs and 18% of cases contained lexical words (besides grammatical words).

Three word repetitions

The number of three word repetitions was very low in our corpus (only 25 repetitions contain 3 words). From these repetitions 32 % contained lexical words. In most of the cases (80%) some pauses (filled or silent) occurred between the constituents of the repetition.

3.1.1. Prosodic parameters in word repetition

The prosodic parameters investigated in the word repetitions were F0 patterns, vowel durations as well as filled and silent pause locations. The vowel durations and the F0 slopes were measured on the last vowels of the repeated words or of the repeated word groups. This way, for example, when a repetition contained 3 words, it was then the last vowel of the third word in both parts of the word groups, which was considered.

F0 pattern

As illustrated in Figure 1, most of the F0 patterns were flat on both last words of the repeated sections. The F0 movement was slightly smaller on the first word than on the second one on which the number of low-low and high-high movements were higher than on the first word. In French the syntactic junctures are marked with an F0 movement containing a clearly downward (at sentence final position) or upward (at non-final position) patterns. However, speech repetition contains a different pattern with a very slight (flat) or a moderate (mid-low) movement of the F0. Such a flat F0 slope is a prosodic cue of an unfinished speech sequence.

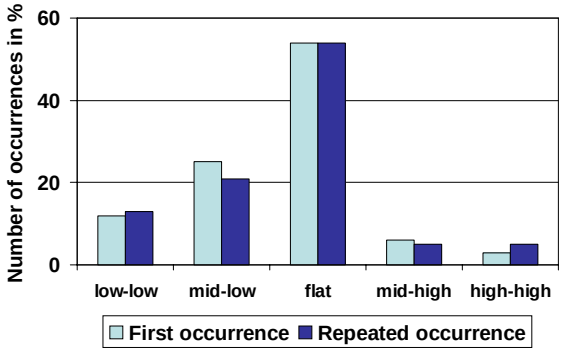

Figure 1: F0 patterns in word repetitions. Number of occurrences as %

Vowel duration in the last syllable

Vowel durations were measured only on the last syllable since the lengthening of the last syllable was one of the prosodic cues of the prosodic boundary. A long final word syllable, accompanied with a clear F0 movement, signaled the presence of a prosodic parsing.

As illustrated in Figure 2, in word repetitions, the first word occurrence contained a stronger vowel lengthening than the repeated word. The hesitation, when present in word repetitions, was expressed by a longer last syllable (and a flat F0 pattern). On the other hand, almost 80% of all the vowel lengths measured on the last part of the repetition sequence were shorter or equal to 100 ms. A short vowel length in this position clearly indicated the absence of a prosodic boundary.

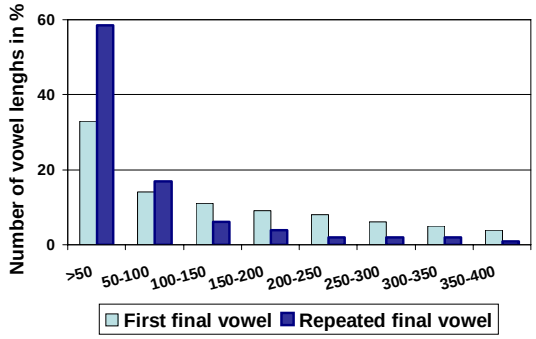

Figure 2: Vowel duration in one word repetitions

Filled pauses in word repetition

In word repetitions the use of filled pauses concerned only 17% of messages which corresponded altogether to 179 filled pauses. Of these filled pauses 26% were placed before the repeated words, only 9% after the repeated words and the remaining 65% between the words of the repetition. Multiple filled pauses in a repetition rarely occurred (only 17 cases that is 9%).

A filled pause can be stuck to the previous word (26%) stuck to the following word (33%) stuck to the previous and following words (19%) but most of the time (75% of the cases) filled pauses were separated from the surrounding words (previous or following or both) by silent pauses.

Filled pauses when non-separated by a silent pause from the surrounding words, followed the final consonants but also the final vowels of the previous word. When a consonant ended a word, the filled pause was perceived as a lengthened schwa vowel and when a vowel ended a word, its timber was "neutralized" into a schwa one. The mean value of the filled pause length in speech repetition was 355 ms with a large standard deviation (251 ms).

Silent pauses in word repetition

70% of word repetitions contained silent pauses. The silent pause mean value was 190 ms and its standard deviation was very large (289 ms). Word repetitions containing silent pauses started with a silent pause in 24% of cases; they ended with a silent pause in 21% of cases and contained silent pauses inside the repetition in 55% of cases. 44% of the word repetitions contained more than one silent pause. 60% of all pauses present in word repetitions were short pauses. The amount of intermediate (mid) pauses reached 26% and of long pauses 14%. The number of pauses was correlated with the number of repeated words: the higher the number of words in a repetition was and the higher the number of pauses encountered in it. In three word repetitions 80% of all the pauses occurred between the words of the repeated sequences. In our corpus the highest number of silent pauses encountered in three words repetition was 7.

3.2. Pauses

Speech can be perceived as highly disfluent even when no "loud" disfluency cues are present in it. That happens when the speech signal contains a high number of pauses, especially when the pause locations are not congruent with the syntactic parsing. Pauses are vital in speech production as the speaker must breathe but they are also necessary for the listener for they provide time to decode the speech stream. In fluent speech pauses are situated on syntactic boundaries signalling the syntactic parsing of the speech or are used for stylistic purposes to enhance the word meaning. However, from the data studied, it appeared that pause occurrences in the spontaneous speech did not always follow syntactic parsing.

3.2.1. Silent pauses

12719 speech internal pauses were detected in our data. On average a pause followed every 4th word.

As previously explained, pauses were grouped into 3 categories: short, intermediate (mid) and long pauses. The quantity of **short pauses** (shorter or equal than 150 ms) was **54%** (6920), **21%** (2684) of all the pauses detected were pauses having an **intermediate length** (from 150 to 250 ms) and **24%** (3115) of all speech internal pauses were **long pauses**, longer than 250 ms.

Figure 3 illustrates the number of pauses as a function of the number of words preceding the pause. It appeared for all three categories of pauses that their location after one single word accounted for the highest number of occurrences. Such occurrences of silent pauses suggested that speakers used them as a talk preparation gap and they were not necessarily situated on syntactic boundaries. This was confirmed by F0 pattern values measured before the pauses and reported in Figure 3. As previously recalled, in French the F0 pattern on prosodic boundaries has a clearly downward or upward movement. Yet a great amount of F0 slopes, measured at the vicinity of pauses, is a flat one (with a very slight F0 movement). Therefore it can be supposed that pauses preceded by a low number of words and by a last syllable having flat F0, occur where hesitation is present.

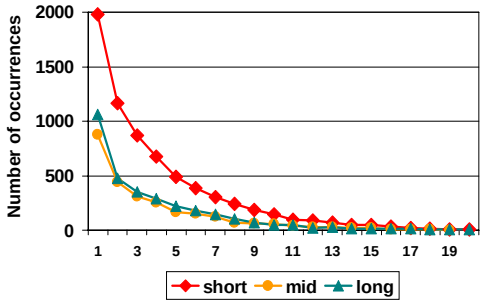

Figure 3: Number of pauses as a function of the number of preceding words

As seen in Figure 4, there were less short pauses when the F0 movement had an upward direction, nonetheless they were slightly more frequent when the F0 movement had a downwards direction. This was due to the fact that a clearly upward movement of the F0 encoded a prosodic parsing which required most of the time long or intermediate pauses.

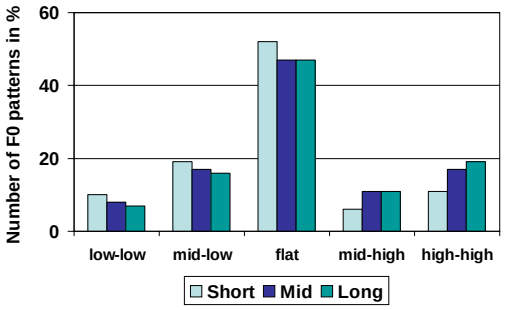

Figure 4: F0 patterns as a function of the following silence pause length.

Figure 5 represents mean vowel lengths as a function of the following pause and the F0 pattern. There was a positive correlation between the pause length and the vowel duration measured in the syllable before the pause. That means that the vowel length was longer when the word was followed by long pauses than when it was followed by intermediate or short pauses. However, when the F0 pattern was high, then the vowel duration was lengthened no matter how long the following pause was: in this place an occurrence of a prosodic boundary could be supposed.

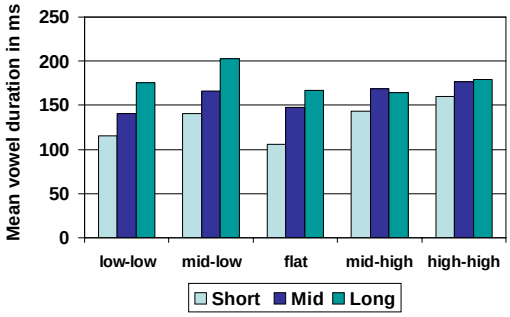

Figure 5: Vowel length when followed by a pause as a function of the F0 pattern and the silent pause length.

In spontaneous speech, even syntactically strongly related words can be separated by pauses. In fact, in our corpus 24% of all pauses occurred between words with strong syntactic links such as pronouns + verbs (je suis – I am) or articles and nouns (la chaise – the chair). Among these pauses 70% were short pauses. In these cases the dominant F0 pattern (60% of cases) was a flat one. Nevertheless the downward pattern was also rather frequent, with 19% of mid-low patterns and 11% of low-low patterns. This was probably due to the fact that grammatical words in French are often uttered with falling F0. One can speculate about the maintenance of this typical F0 pattern of function words despite pause occurrences.

3.2.2. Filled pauses

The number of filled pauses in our corpus was 2871. The filled pauses could be separated from the surrounding words by a silent pause (in 32% of cases). They could be attached to the preceding word (in 68%) as a very long schwa like vowel uttered after a final consonant. They could also be perceived as a neutralized part of the preceding vowel timber when attached to a previous last vowel of a word.

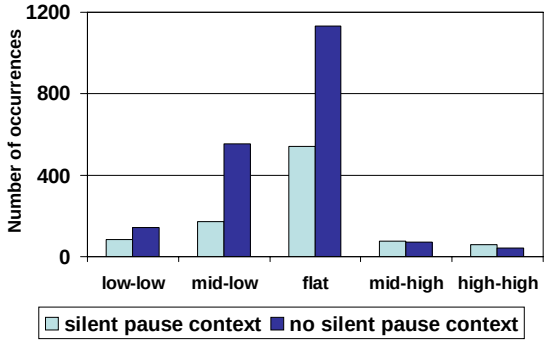

Figure 6: F0 pattern in filled pauses.

The prosodic cues of the filled pauses were very robust. The main cue was its duration (the mean duration was equal to 350 ms) and its F0 movement which was mainly flat. Figure 6 illustrates the prosodic cues of filled pauses followed or not by silent pauses. Although a lot of their F0 patterns were flat, the downward F0 pattern, especially a moderate one (mid-low), was also quite frequent.

3.3. Speech repairs

Speech repair can be a false start containing an interrupted pronunciation of a word or a correction of a wrong word. A simple correction is used mainly for grammatical words when a word has a counterpart considered as the correct one in a given context, such as articles which can be definite or indefinite or masculine or feminine. A lexical word is seldom corrected by adding simply the right word. In fact it is either interrupted or it is corrected by lexical means: such as "la chaise, pardon, la table" (the chair / I mean the table).

Only a very low number of word pronunciations (204) were interrupted in our data base. Speech repairs carried out as a simple word repetition were used only for grammatical words and their number was low (180 cases found in the data). When a speech repair occurred then most of the time (80% of the cases in our data) it was separated by a silent pause. These pauses were in 40% of cases long pauses, in 36% of cases short pauses and in 23% of cases intermediate pauses.

3.4. Other spontaneous speech events

Beside the spontaneous speech events previously discussed in this study, there were others that could be characterized by their prosodic parameters. For example, in spontaneous French, the word "**quoi**" (what) is frequently used by speakers as a loud full stop. In fact, this word had no other role here that to bring down the F0 when its value was too high for a sentence final position. Their prosodic characteristics were therefore very typical. Its prosodic characteristic was a falling F0 (in 80% of cases studied here) and short vowel duration (in 75% of cases shorter than 100 ms).

4. Tests

Prosodic parameters should be used to confirm or refute the solution of the recognition system and to signal disfluency cues (hesitation, filled pauses…). As a preliminary experiment, an assessment of the observations discussed above was carried out on the same corpus and under the same conditions as the statistical study: the phonetic forms of the words were aligned with the speech signal and no speech recognition was carried out.

In order to test how efficient prosodic parameters could be in detecting disfluencies in spontaneous speech, prosodic rules were formulated and their impact tested. The formulation of the prosodic rules was based on the observations of normalized histograms representing F0 slope values and the vowel duration values (see Fig. 7). The slope categories were the same as in the previous part of the study (LowLow, MidLow, Flat, MidHigh, HighHigh) but their values were normalised separately on each recording.

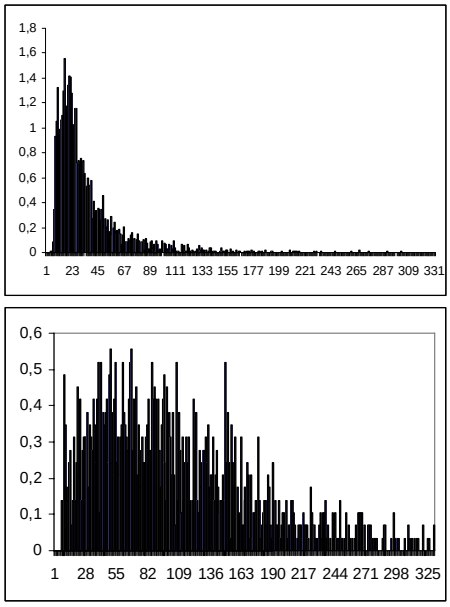

Figure 7: Normalized histograms for word last vowel durations in ms (above) and filled pause durations in ms (under) when the slope is flat.

Tests were conducted to verify how general simple expert-defined rules were and how efficiently they could detect disfluencies in speech. Prosodic rules were used to confirm or not filled pauses. Moreover a flag indicating a hesitation was positioned when the values of the prosodic parameters were considered as conveying a hesitation A hesitation was automatically detected as present when the final syllable was very long (followed or not by a pause) and when it was very short and followed by a pause. In the first case the hesitation was expressed by the syllable length and in the second by the

occurrence of the pause. The correctness of the hesitation flag positions was checked manually, by listening to the speech signal.

The following evaluations were carried out:

Filled pauses: 60% of filled pauses present in the alignments were validated as filled pauses using prosodic parameters about. When the occurrence of a silent pause preceding the filled pause was taken into account, the validation of filled pauses reached 72%.

Hesitation: 33% of the last syllables of content words and 11% of function words were detected as conveying a hesitation.

The accuracy of the hesitation flag positions was verified manually on 10% of our data. It was found that 83% of flags were correctly set. As the transcription of the corpus did not contain information about hesitation (other than filled pauses) therefore it was impossible to estimate the number of missed hesitations.

Word repetitions: No typical prosodic cues were observed in this study for word repetitions. The only prosodic cue detected was hesitation conveyed most of the time by the last syllable of the fist part of the repeated word sequence.

This preliminary evaluation was very promising. However further work is to be conducted using speech recognition results and also other types of data such as data from man-machine speech driven "dialogues". On the other hand appropriate modelling procedure is to be used in order to refine the disfluency decision thresholds.

5. Discussion

The present study analyses the prosodic cues of disfluencies in French spontaneous speech. A very long vowel duration in a last syllable of a word is a reliable disfluency cue. However a flat F0 slope and a short vowel duration in a last syllable of a word followed by a silent or a filled pause can also be considered as a strong prosodic indicator of disfluencies. A small number of words separated by pauses can be another indicator of places in speech where hesitations or unease occur. The filled pauses are characterized in prosodic terms with a long duration and a flat or a slightly downwards F0 movement.

When expert defined simple prosodic rules are used to test how efficient prosodic parameters in disfluency detection are, it appears that they are very reliable in filled pause detection and efficient in hesitation detection which is often present in speech repetitions. Thus, prosodic parameters can be used in speech recognition to yield confidence indication about the occurrence of disfluencies.

6. References

[1] Asp Annika & Decker Anna 2001. Designing with speech acts to elude disfluency in human-comuter dialogue systems, Working Papers 49, 2-5 Lund University.

[2] Arnold Jennifer E., Fagnano Maria & Tanenhaus Michael Disfluencies Signal Theee, Um, New Information, Journal of Psycholinghistic Research, Vol. 32, No. 1, January 2003

[3] Bailey, Karll G.D., & Ferreira, Fernanda. 2001. Do Non-Word disfluencies Affect Syntactif Parsing?, DISS'01, Edinburgh, Scotland, UK pp. 61-64

[4] Baron Don, Shriberg Elizabeth & Stolcke Andreas 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues, Icslp 2002.

[5] Bartkova, Katarina & Sorin Cristel, 1987. A model of segmental duration for speech synthesis in French, Speech Communication 6, pp 245-260.

[6] Bortfeld Heather, Leon Silvia D., Bloom Jonathan E., Schober Michael F. & Brennan Susan E. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender, Language and speech, 2001, 44 (2), pp. 123-147.

[7] Honal Mathais & Schultz, Tanja 2003. Correction of Disfluencies in spontaneous speech using a noisy-channel approach, EuroSpeech 2003;

[8] Hutchinson Ben & Pereira Cécile 2001. Um, One Large Pizza. A Preliminary Study of Disfluency Modelling for Improving ASR, DISS'01, Edinburgh, Scotland, UK, pp 77-80;

[9] Martin Philippe 1981, Pour une théorie de l'intonation, *L'intonation de l'acoustique à la sémantique*, Klincksieck Paris pp. 234-271.

[10] Rose R.C & Riccardi G.1999. Modeling Disfluency and Background events in ASR for a natural language understanding task, ICASSP 1999;

[11] Shriberg Elizabeth, Batte Rebecca & Stlcke Andreas 1997, A prosody-only decision-tree model for disfluency detection, Eursopeech 1997.