Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop.
5–8 September 2003, Göteborg University, Sweden, pp. N–N.

A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models

M. Adda-Decker, B. Habert, C. Barras, G. Adda, Ph. Boula de Mareuil & P. Paroubek

LIMSI-CNRS Orsay France

Abstract

The aim of this study is to elaborate a disfluent speech model by comparing different types of audio transcripts. The study makes use of 10 hours of French radio interview archives, involving journalists and personalities from political or civil society. A first type of transcripts is press-oriented where most disfluencies are discarded. For 10% of the corpus, we produced exact audio transcripts: all audible phenomena and overlapping speech segments are transcribed manually. In these transcripts about 14% of the words correspond to disfluencies and discourse markers. The audio corpus has then been transcribed using the LIMSI speech recognizer . With 8% of the corpus the disfluency words explain 12% of the overall error rate. This shows that disfluencies have no major effect on neighboring speech segments. Restarts are the most error prone, with a 36.9% within class error rate.

1. Introduction

Within the Human-Machine Communication department at LIMSI, we experiment with combining skills and techniques in audio document transcription and in text processing, in order to improve both domains.

A first step consists in processing texts and audio documents belonging to the same topic, as such 'sibling' resources become more and more within reach (for instance broadcast news and newspaper articles about the same event). Using such resources to improve speech transcription and structuring represents a first direction. These texts can simply share the topic of the audio documents. In the present study, however, they consist in relatively close transcriptions. For instance, these texts can help to define the language model (topic-centered models, specific lexica). A better knowledge of spontaneous speech then becomes necessary to also improve modeling along the relatively topic-independent dimension of spontaneous speech phenomena.

A second direction consists in enhancing speech transcriptions so as to input them to taggers, parsers and indexing tools. As these softwares may require punctuation marks to facilitate sentence, clauses and phrases identification, linguistic models of punctuation and acoustic hints such as break length, inspirations, intonation can be used to complete the transcriptions in this area. On the other hand, disfluencies (DFs in the sequel) must be edited to get text-like input or to improve the readability of an automatic transcription.

This study makes use of 10 hours of French radio archives, recorded about 10 years ago (we will refer to this corpus as the *archive corpus*). In each one hour show a major personality from either political or civil society (e.g. nonprofit humanitarian organizations...) undergoes a detailed questioning by quite a few journalists. The setting of the chosen interviews favors the production of disfluencies. One of the reporters acts as chairman. He monitors the repartition of time between the reporters and the before-hand chosen topics; he often interrupts the interviewee or the reporter: overlaps are frequent. Each reporter has a 'slot' for prepared questions on a given topic; even the interviewee's answers are not entirely spontaneous: most questions are obvious ones and trigger prepared answers. The interviewee has often been coached before the show. Therefore, in our data, speech is neither entirely spontaneous nor totally planned. It is better described as constrained. These constraints yield numerous disfluencies. Only part of them reveal information about the planning problems of the speaker [5]. The rest of them resorts to the 'struggle for speech' between interviewer and interviewee, or between interviewers, even though reporters probably do not "jump in" at random locations [6].

For each show we have both the audio data and press-oriented transcripts. These press-oriented transcripts (TPress henceforth) are intended to be rather close to the audio (as quotations are being extracted from them for other media) while lying somewhere in between written text and exact transcript: they stick to implicit conventions for speech rendering. As a matter of fact most disfluencies and linguistic errors have been discarded or edited. We produced as well an exact audio transcription (TExact) for 10% of the data: all audible phenomena in particular disfluencies (for spontaneous speech modeling studies) and overlapping speech are manually transcribed. We relyed on standard automatic transcripts (TReco) to tune the exact transcripts: it is indeed easy to 'miss' some disfluencies and to inconsciously edit them.

The comparison of the TPress, TExact and TReco transcripts allows us to investigate the following questions: what is the overall proportion of DFs observed? within DFs, what is the repartition between the different types? is this repartition correlated with sociological features of the speakers or with competition for "foreground" speech? Are the different DF classes more or less error prone? Are they difficult to take into account using conventional word N-grams?

2. Spontaneous speech annotations

Spontaneous speech, with its hesitation phenomena, repetition of function words and other false starts, has hosted a great deal of interest from several French teams. Morel and Danon-Boileau [4] who especially studied intonation (in particular that of parentheticals), addressed these "little words" typical of spoken language, which they call "ligators": e.g. *quoi, ben, enfin* ("well"), *donc, alors* ("so"), *genre, style* ("kind of").

The GARS, in Aix-en-Provence [2] worked for years on the problems raised by transcribing speech. The choices, which written representations assume, with a grammatical exploitation of spoken corpora in view, are a trade-off between faithfulness and legibility: a transcription in standard orthography is given, without "faking" — no transcription under the morpheme level is foreseen. No punctuation marks are specified since they yield an a priori segmentation into phrases or sentences, which prejudges the analysis. Another project, PFC (Phonologie du Français contemporain) [3], in a socio-phonological framework aiming at covering a vast geographical area, recently started to take prosody into consideration. The objective is to align the spoken data with written texts as easily as possible: hence the choice for an orthographic transcription which includes standard punctuation marks. Background manifestations such as *hum* are ignored and not transcribed. Hesitations are transcribed by *euh*, even when it is difficult to distinguish them from the pronunciation of a schwa.

The annotations adopted in our work partially rely on the LDC's metadata annotation guidelines [7] used for the Rich Transcription evaluations conducted by NIST (http://www.nist.gov/speech/tests/rt/rt2003/). These guidelines aim at producing maximally readable transcripts: "[...] annotators will identify fillers, depods (the deletable portion of an edit disfluency), and SUs ('semantic units'). Transcripts [...] can be cleaned up for readability; for instance, depods and fillers must be removed and each SU presented as a separate line within the transcripts". We chose these guidelines because they are consistent with our own objectives and represent the current result of a vast discussion.

SUs are coarsely defined as 'units within the discourse that function to express a complete thought or idea on the part of the speaker,' with a pragmatic aim in mind: '[...] the goal of SU labelling is to improve transcript readability by creating a transcript in which information is presented in small, structured, coherent chunks rather than long turns or stories.'

Fillers are divided between filler words (FW: like *um*), discourse markers (DM in the sequel: 'a word or a phrase that functions primarily as a structuring unit of spoken language'), explicit editing terms (EET: 'overt statement from the speaker recognizing the existence of disfluency'), asides (AS: 'the speaker utters a short comment on a new topic then returns to the main topic being discussed'), parentheticals (PA: 'the remark is on the same topic as the larger utterance'). Edit disfluencies (ED) are divided between repetitions (RP in the sequel), revisions (RV in the sequel), restarts (RS: 'the corrected portion that replaces the depod modifies its meaning'), and complex disfluencies.

For the annotation of the archive corpus, we decided to follow as much as possible the LDC guidelines and to adapt them to French with some simplifications. We marked PA and AS in the exact transcriptions, but we do not comment on them. We merged RV and RS under the heading RS, as it is not always easy to assess the intended modification of meaning between the depod and what follows it.

3. Corpus and transcriptions

3.1. Corpus and exact transcription

In the sequel, each speaker is given an ID (from 1 to 20), followed by letters refering to some of his/her sociological features as shown in the table below. Letters are necessarily one of **J** or **I.** If not more specified a speaker is by default a French adult man. There is just one woman among the interviewees. Interviewees are by default politicians.

code	meaning	#spk
J	journalist	9
C	chairman	1
e	English native	1
r	region. accent	1
f	francophone	2

code	meaning	#spk
I	interviewee	11
w	woman	1
o	elderly	1
c	Civil society	2

One of them is an English native speaker, two persons are French native speakers from African francophone countries.

For our study, an exact audio transcription has been produced manually on 10% of the corpus: 2 excerpts of approximately 3 minutes, selected randomly in each show, have been split in SUs and all disfluencies have been explicited and annotated according to the previously detailed guidelines.

The range of words per SU is between 8.6 and 20.8 (median: 12.8, mean: 13.7). Median and mean are greater for interviewees than for journalists (median: 13.9/11.6, mean: 14.9/12.2). Interviewees make longer SUs than journalists.

In order to characterize the speakers according to their disfluencies, Correspondence Analysis (CA) was being used for features DM, RP, RS, FW (see Figure 1). CA provides the best fit, in the least squares sense, relative to the chi-squared distance, to both the speaker points and the disfluencies points. It yields a sequence of orthogonal axes. We show the two first axes. The projection of the points shows no obvious clustering, neither of journalists, nor of interviewees. The first axis opposes dominance of RS (16-Io, 19-If, 20-Ie) to dominance of DM (12-I, 10-Iwc). The second axes contrasts the association of FW and RP (1-JC, 17-Ir, 3-J) to the important use of DM (6-J, 13-I).

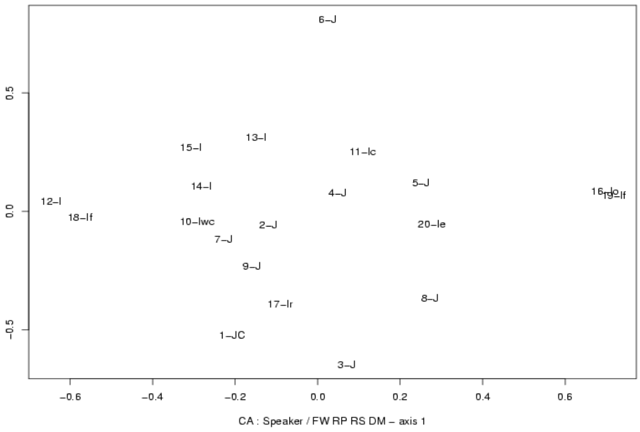

Figure 1: Proportions of DFs (FW,RP,RS) in the TExact transcripts for each speaker.

Even though CA provides no clear opposition between journalists and interviewees as such, the balance between RP and RS seems to correspond to different 'choices'.

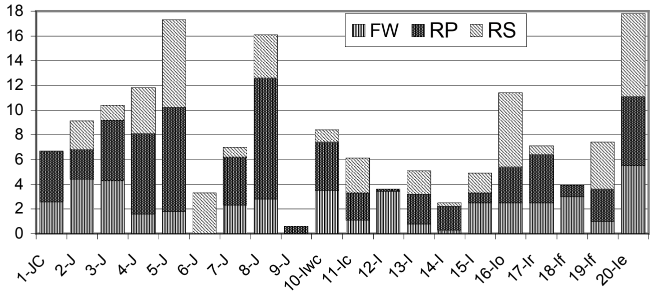

Figure 2: Proportions of DFs (FW,RP,RS) in the TExact transcripts for each speaker.

As shown in Figure 2, within DFs, for 13 speakers (7 journalists: 1-JC, 3-J, 4-J, 5-J, 7-J, 8-J 9-J; 6 interviewees: 10-Iwc, 12-I, 13-I, 14-I, 17-Ir, 18-If), the proportion of RP is greater than the proportion of RS. 5 interviewees (11-Ic, 15-I, 16-Io, 18-lf, 20-Ie) and one journalist (6-J) show the opposite situation. The use of RP is clearly dominant within journalists, possibly because of the difficulties journalists meet while trying to interrupt interviewees. On the opposite, RS have the first role for half of the interviewees: in spite of journalists interviewees seem to have real opportunities in tuning their words.

3.2. Automatic audio transcription

The audio corpus has been transcribed using the LIMSI speech recognizer resulting in the TReco transcripts.

3.2.1. Recognition system description

The LIMSI standard broadcast news transcription system for French [1], was used for transcribing the one-hour subset of the corpus. The acoustic models were trained on about 100 hours of French broadcast news data; they consist in context-dependent models of 33 French phonemes, plus 3 generic models for silence, filler words and breath noises. The standard language model (LM) is an interpolation of 4-gram back off language models trained on different data sets. Three different sources were used: press-oriented transcriptions of various broadcast shows (48M words), exact transcriptions of broadcast news (BN) data, mainly radio shows (0.95 M words) and newspapers texts (311M words). The lexicon contains 65k words, chosen for optimizing the coverage of broadcast news development data (very different in date and source from the archive corpus). The pronunciations are derived from grapheme-to-phoneme rules and manually checked. The system runs at about 10 times real-time on a standard PC.

Using the press-quality transcriptions provided with the corpus (about 580k words), an "informed" LM was designed by interpolation with the standard n-gram LM; the lexicon contains only the 26k most frequent words from the standard sources, together with all the 19k words contained in the press-oriented transcripts, resulting in a 30k words lexicon.

3.2.2. Standard and informed recognition results

Performance of automatic speech transcription was evaluated using NIST sclite tool, by counting the percentage of word differences relative to the TExact transcription. Disfluencies were tagged in the reference as optional words, i.e. no error was counted if a filled pause or a word involved in a repetition or a revision was ignored by the system. Most of the overlapping speech, where the speakers clearly speak in synchrony, has been discarded from the evaluation. However a non-negligible amount of overlapping speech remains as speech on background noise: the 2nd speaker uttered just one or two words over a sentence of the 1st speaker (back-channel), the background speech is not intelligible. These speech on speech background noise segments are the most error prone.

Using the standard French transcription system, an average word error rate of 24% could be measured. The relatively high word error rate on this data should be compared with two other figures: our standard word error rate on other BN data in French (about 20%) and our current result on the last RT03 evaluation on American English broadcast news data (11.7%). From this comparison, we may expect improvements from:

1. working on French specificities: recently, we have focussed our work on American English. If similar efforts are done for French [1], we may expect to reduce the gap between French and English systems to less than 5%. In particular, we may increase the quantity of data used to build the language model: we currently use 5 times less data for BN French than for BN English.
2. developing specific acoustic models for the archive corpus, to reduce the gap with our standard word error rate in French BN.

In a second transcription experiment, the informed LM was used: the resulting word error rate is 14.5%, a 40% relative reduction as compared to the 24% obtained with the standard system. One purpose of our experiments with informed transcriptions was to test if accurate transcriptions can be obtained starting with fast-to-produce press-quality transcriptions. The high word error rate shows that simply feeding the press-quality transcripts into the language model is not enough for producing high quality transcripts. By contrast, it also shows errors, which mainly stem from acoustic problems. Per-speaker results are given Table 1 and show a large inter-speaker variability.

Table 1: word error rates of the standard system (*WER-S)* and informed system (*WER-I*).

Journalist	WER-S	WER-I	Interviewee	WER-S	WER-I
1-CJ	33.0	22.7	10-Iwc	24.2	14.2
2-J	19.7	13.3	11-Ic	25.5	13.6
3-J	16.9	10.1	12-I	17.4	4.9
4-J	23.7	11.2	13-I	19.8	10.3
5-J	25.8	17.1	14-I	16.6	8.6
6-J	18.8	6.0	15-I	16.7	9.8
7-J	36.2	23.8	16-Io	35.0	21.2
8-J	24.6	23.8	17-Ir	27.8	16.7
9-J	14.0	3.0	18-If	28.4	15.5
All (I+J)	**24.0**	**14.5**	19-If	32.7	24.4
			20-Ie	28.7	22.5

4. Comparison of manual transcripts

The press-oriented transcripts are fairly close to the audio data. To get an idea of the differences between both TPress and TExact versions, sclite is used again, with, as a reference, the TExact version where all disfluencies have been filtered out. Word difference rate amounts to 9%. Disfluencies are obviously not the only reason of differences between the two versions.

4.1. TExact *vs* TPress: deletions, insertions substitutions

A more detailed study of differences between both versions showed the following:

Deletions other than disfluencies occur and are mainly due to omitted parentheticals, asides or DM sequences. Example:

~~oui mais je pense qu'aujourd'hui si vous voulez~~ l'économie du monde a commencé à changer dans les années soixante.

In press transcripts overlapping speech is considered as two consecutive flows whenever possible: this generates a significant part of word **insertions** and highlights the problem of overlapping speech in this kind of data. Other more French specific phenomena entail insertions: reductions like "*y a*", "*c'est pas*", which correspond to the effectively produced speech, are transcribed in correct written French as "*il y a*", "*ce n'est pas*". As this kind of reductions appear (as in other languages), on the most frequent word sequences, their global impact on insertion rates is significant.

Substitutions are also often due to reduction phenomena: the pronounced word "*ça*" (engl. "*it*", reduced form) is most often transcribed using its canonical form "*cela*" . Other reasons are verb tense or mode (*voulais* vs *voudrais*), gender (*nous l'avons établi* vs *nous les avons établis*), interrogative forms (*est-ce qu'on doit* vs *doit-on*), numbers (*un milliard huit* vs *un virgule huit milliards*). Very few differences are due to human errors (e.g. date of an historical event).

4.2. Typology of observed disfluencies

In the following we consider the 3 main disfluency types: filler words (FW), repetitions (RP) and restarts (RS) including revisions here. The FW class contains a single element *euh*.

The major part of RPs are of the simplest form: two consecutive monosyllabic words. Good candidates for repetitions are articles, pronouns, prepositions, adverbs. The most observed items are: *le, de, un, à, et, qui, que, les, très, pas*. Of course more complex repetition structures are observed (*~~beaucoup de,~~ beaucoup de* ; *~~peut-être~~ alors peut-être* ; *~~et et et et et et le plus~~ et le plus ...*), but they account only for a low percentage of repetitions.

The RS class is the most heterogeneous one. Revisions can simply be due to an anticipated erroneous form or gender determination (*~~pour le~~ pour l'événement*), which needs correction. Beyond this simple category, any phrase can be revised or restarted and no synthetic overview can be given.

About 8% of the TExact words are in the three FW, RP or RS classes (with 2.5, 3.2, 2.3% of the words respectively).

In addition 6.3% of the words correspond to discourse markers (DM). DMs are not really disfluencies, but specific events of spontaneous speech. Their role is more or less to introduce speech or to glue speech sequences together. They seem particularly useful in the struggle for speech situation. A limited number of words are generally observed as DM: *alors, et, mais, donc, bon, voilà, oui, hein*. However each speaker may have its own preferences and habits of DMs.

5. Disfluencies and recognition errors

The TReco form of the corpus contains 9400 words (approx. 1 hour of speech) and 1365 errors (14.5%). We are interested in measuring the contribution of disfluencies to the overall error rate. Table 2 shows the major error sources, starting with the introduced disfluency classes and the discourse marker class. Beyond disfluencies and spontaneous speech specific words like DMs, pronunciation reductions (PR) on common words and word sequences are a serious source of errors. Whereas disfluencies alone account for about 12.5% of the observed errors, DMs produce 8.2% of errors. A more important contribution of 25.1% comes from the reduced pronunciations.

Table 2: Number of errors observed in different classes. The first classes correspond to disfluencies. The last class focuses on pronunciation reductions, fast and badly articulated speech (PR). For each class its contribution to the overall error rate is given.

Class	*#errors*	*% overall error*
FW+RP+RS	171	12.5%
FW+RP+RS+DM	283	20.7%
PR	347	25.1%

It is also interesting to know whether disfluencies are significantly more error prone than other words.

Table 3: Within class and overall error rates for the main DF classes.

Class	*#errors/#total*	*% errors in class*	*%overall error*
FW	45 / 231	19.5%	3.0%
RP	46 / 300	15.3%	3.0%
RS	80 / 217	36.9%	6.5%
DM	112 / 593	19.3%	8.2%

Table 3 shows for each class the number of errors and the total number of words observed in this class and the corresponding within class error rate.

Whereas all the class-specific error rates are above the average corpus error rate, some classes are seen to be more difficult to handle than others: 36.9% errors for RS vs. 15.3% for RP. Significant differences are also observed between speakers. Among the interviewees a non-native person produces half of all the errors on repetitions (23 errors). By just excluding this speaker from the counts, the repetition error rate falls to 8.8%, which is far less than the average error rate (13.8% without the non-native speaker).

6. Discussion

In this study we have compared different types of audio transcripts with, as objectives, a better modeling of spontaneous speech specifities and their appropriate rendering in audio transcripts.

The comparison of press-oriented and exact audio transcripts showed that disfluencies explain only about half of the observed differences. Discourse markers, parentheticals, rewording and overlapping speech transcriptions are the main factors for the additional differences. Whereas many disfluencies may simply be filtered out in the transcriptions, others carry some information: hesitations may indicate syntactic disfluencies and keeping some marks increase readability and acceptability.

Concerning automatic transcription we investigated the impact of disfluencies on word error rates. With 8% of the corpus the disfluency words explain 12% of the overall error rate. This shows that disfluencies have no major effect on neighboring speech segments. Restarts are the most error prone, with a 36.9% within class error rate. However dealing with restart phenomena on a simple lexical level appears to be insufficient: including morpho-syntactic information may provide a useful modeling level here. If overlapping speech is held out, reduced pronunciations appear to be the major error source: results may be significantly improved if these phenomena are better taken into account, in both the pronunciation dictionary and the acoustic models.

Another aim concerns the automatic production of exact audio transcipts using press-oriented corpora. Even if improvements are still in reach using standard developments, more spontaneous speech specific research seems required given the relatively high error rates observed with informed language models.

7. References

[1] Adda-Decker Martine, Gilles Adda, Jean-Luc Gauvain, Lori Lamel, *Large vocabulary speech recognition in French* , Proc. IEEE ICASSP'99, I, pp.45-48, Phoenix, AZ, March 1999.

[2] Blanche-Benveniste Claire (1990), *Le français parlé, études grammaticales*, Éditions du CNRS, Paris.

[3] Delais-Roussarie, Elisabeth & Jacques Durand (2003), *Corpus et variation en phonologie du français:méthodes et analyses*, Presses Universitaires du Mirail, Toulouse.

[4] Morel Marie Annick & Laurent Danon-Boileau (1998), *Grammaire de l'intonation. L'exemple du francais*, Éditions Ophrys, Paris.

[5] Plauche, M. and E. Shriberg, (1999). *Data-Driven Subclassification of Disfluent Repetitions Based on Prosodic Features*. Proc. International Congress of Phonetic Sciences, vol. 2, pp. 1513-1516, San Francisco.

[6] Shriberg, E., A. Stolcke, A. & D. Baron (2001). *Can Prosody Aid the Automatic Processing of Multi-Party Meetings?* Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, pp. 139-146, Red Bank, NJ.

[7] Strassel Stephanie, Simple Metadata Annotation Specification Version 5.0 – May 14, 2003, http://www.ldc.upenn.edu/Projects/MDE/