A CRF-based Approach to Automatic Disfluency Detection in a French Call-Centre Corpus

Camille Dutrey[1, 2, 3, 4], Chloé Clavel[5], Sophie Rosset[2], Ioana Vasilescu[2], Martine Adda-Decker[2, 3]

[1]EDF R&D, 92 141 Clamart, France
[2]LIMSI-CNRS, 91 403 Orsay, France
[3]LPP, Université Paris III – Sorbonne Nouvelle, 75005 Paris, France
[4]Université Paris Sud, 91405 Orsay, France
[5]Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, 75014 Paris, France
{dutrey,rosset,ioana,madda}@limsi.fr, chloe.clavel@telecom-paristech.fr

10.21437/Interspeech.2014-601

Abstract

In this paper, we present a Conditional Random Field based approach for automatic detection of edit disfluencies in a conversational telephone corpus in French. We define disfluency patterns using both linguistic and acoustic features to perform disfluency detection. Two related tasks are considered : the first task aims at detecting the disfluent speech portion proper or reparandum, i.e. the portion to be removed if we want to improve the readability of transcribed data ; in the second task, we aim at identifying also the corrected portion or repair which can be useful in follow-up discourse and dialogue analyses or in opinion mining. For these two tasks, we present comparative results as a function of the involved type of features (acoustic and/or linguistic). Generally speaking, best results are obtained by CRF models combining both acoustic and linguistic features.
Index Terms : disfluencies, conditional random fields, conversational speech, spontaneous speech.

1. Introduction

The last decade witnessed a growing interest in *speech analytics* of various types of call-centre data for marketing applications. This particular focus on spontaneous, interactive telephone data brings known scientific challenges to the foreground. A key issue is the development of information extraction systems dealing with spontaneous speech features at different linguistic levels : acoustic-phonetic, lexical, syntactic, semantic, dialogic. For decades, specific phenomena of spontaneous speech were regarded as detrimental to the quality of discourse and to its comprehension. In particular, utterance breaks, with elements "disrupting" the syntagmatic progress of the linguistic message, are often considered as signals of a spoken message under construction – such as various drafts prior to a final written text. The study of spontaneous oral language has long been neglected in linguistic studies. However, during the last decades, more focus was put on spontaneous speech phenomena [1]. Recent studies considered the lexical status of discourse markers and of the larger class of disfluency phenomena. Their importance in conversation structuring and modelling has been underlined, in particular as dialogic cues [2].

The aim of our study is to develop a system able to detect disfluencies in the difficult context of call-centre data. The present approach contributes to an important challenge because analyses are driven in a real industrial context with call-centre conversations provided by the French EDF power supply company. It is motivated by the strategic role of mining of call-centre data for marketing applications. Indeed, the information content of such call-centre conversations is highly valuable to industrials, as it may be used to improve customer knowledge and customer relationship management. In particular, it may contribute to highlight interaction dysfunctions or good practices, poor vs. effective communication strategies between client and agent thus enabling the elaboration of optimised communication strategies for EDF agents. The long run objective of our study is thus to use the developed disfluency detection system to improve the reliability and the efficiency of the text mining methods that are currently used at EDF [3]. In particular, two applications are targeted : improve the transcribed data readability, especially as part of the display interface already developed in the VoxFactory project [4], and help downstream automatic natural language processing modules with the integration or removing of such disfluent events.

The remainder of this paper is organised as follows. In Section 2 we present related work on disfluency detection. Section 3 presents our human/human French call-centre corpus, the VOXFACTORY corpus, and the data used for training and test of our detection system. In Section 4, we present our method of edit disfluency detection using Conditional Random Fields and combined lexical and acoustic features. In Section 5 we present results of our detection system before concluding in Section 6.

2. Disfluencies : from definition to detection

In this study, we focus on edit disfluencies as defined in [1, 5] and adopted by the Linguistic Data Consortium [6] : they are "portions of speech in which a speaker's utterance is not complete and fluent ; instead, the speaker corrects or alters the utterance, or abandons it entirely and starts over". Their structure (see [1, 7]) may be illustrated by the template and a simple example hereafter : [reparandum] * <edit. phase> repair ("I pay [by] * <uh well> by credit card").

Most of the studies on disfluency detection aim at identifying as accurately as possible disfluent areas in order to remove them prior to further speech processing tasks. Systems may make use of acoustic features [8, 9], lexical ones [10] or a combination of both [11, 12]. Many different approaches were tested. [13, 14] used a TAG-based noisy channel approach to model speech repairs and identify edited words. [12] shows

that CRFs outperform other approaches such as Hidden Markov Models or Maximum Entropy. CRFs are also used by [15] (with a post-processing based on Integer Linear Program). State-of-the-art results have been recently obtained for a joint dependency parsing and a disfluency detection task by [16] (using a deterministic transition-based parser) and for reparandum detection task by [17], using Max-Margin Markov Networks and both acoustic and linguistic features (F-score=84.1%). Most of the studies have been conducted on Switchboard (conversational telephone speech in English), which offers a large amount of data [11, 13, 15, 18, 14, 16, 17]. In French, the state of the art is not so provided, be it for available annotated corpora or for detection systems. There are a few automatic detection studies, proposing rule-based models for detection [19], applied to highly specialised domains [20] or in the flow of another principal classification task [21].

As regards removing disfluent areas, most studies exclusively focus on reparandum detection. The perspective proposed here is different. In addition to remove the disfluent area itself, we are also interested in the speaker's efforts to modify his/her statement. Our hypothesis is that the correction area can provide valuable information for dialogue understanding. However, automatic detection and structuring of edit disfluencies remains a challenging task. We propose to explore CRFs to address the challenge of disfluency detection in French data. Their ability to consider large sets of potentially redundant features and to integrate structural dependencies between classes also contribute to the choice of a CRF approach in the current study.

3. A human/human call-centre corpus

We make use of the French VOXFACTORY corpus which has been developed through the eponymous project [3] as a continuation of the Infom@gic–Callsurf project [22]. The French power supply company EDF conducted a recording campaign in a call-center resulting in the VOXFACTORY corpus. This dataset covers a large amount of topics about the company services, e.g. opening contract, technical issues, etc.

3.1. A subcorpus manually annotated in edit disfluencies

A subset of the VOXFACTORY corpus, the VOXDISS set (60 conversations manually transcribed between company agents and individual clients), was manually annotated in edit disfluencies by Vecsys company, a partner of the VOXFACTORY project, with the annotation tool `Transcriber` [23]. The annotation strategy refers to the Linguistic Data Consortium metadata annotation guidelines [6], as described in Section 2. This corpus presents a large variety of disfluency types : repetitions, revisions, restarts, complex disfluencies (see [3] for more details).

3.2. Experimental dataset

Table 1 presents the training, development and test data subsets. To compose each subset, we randomly picked up (in the VOXDISS corpus) conversations respecting an homogeneous distribution over the dialogues' durations. For now, we choose to work on manual transcriptions in order to provide a system that will be independent of the evolution of automatic speech recognition (ASR) systems.

	Train	**Dev**	**Test**
# Calls	48	5	7
Avg. total dur.	9h19	1h01	1h13
Avg. calls dur.	12'04	12'20	11'29
Avg. # speakers	2.1	2.2	3
Avg. # disfl. (per call)	62.54	54.20	37.14
Disfluency density	10.91%	8.60%	7.40%

TABLE 1 – Description of the VOXDISS training, dev. and test data. Disfluency density : rate of words in edit disfluency areas.

4. Automatic detection using CRFs

This section is dedicated to the automatic detection of disfluencies with the CRF method ([24]). We use the CRF implementation provided by `Wapiti` [25], with the `rprop` algorithm as it has been observed that it allows better results on such structured and complex tasks (see for example [26]). Moreover on the development corpus it allows better results. The stopping criterion empirically fixed at 500 iterations maximum, after results obtained on the development corpus. This section is structured as follows : Section 4.1 presents the different tasks that we have implemented for disfluency detection and the corresponding label sets. Section 4.2 describes the specific extraction patterns defined for our model.

4.1. Tasks definition and labels set

The work hypothesis adopted in this study is that detecting disfluencies involves both detecting the global disfluent region and its various elements. In this section, two detection tasks are considered : in **Task-I**, the detection of the disfluent speech portion, including reparandum and editing phase (i.e. the portion to be removed if we want to improve the transcribed data readability) and in **Task-II** the detection of all the elements included in an edit disfluency (reparandum, editing phase and repair). The long run objective is to improve speech analytics according to the dedicated application challenges highlighted in Section 2.

To this purpose, we set up two experiments : in T_1XP_1, the objective is to identify both reparandum (***Rpd***) and editing phase (***EdP***) area as a unique sequence. The associated label is ***Rpd-Edp*** ; in T_1XP_2, we want to distinguish the editing phase and the reparandum and use the two labels Rpd and EdP.

Furthermore, two experiments are carried out in Task-II : in T_2XP_1, which refers to the T_1XP_1, the objective is to add the repair (***Rpr***) detection at the Rpd-Edp region ; in T_2XP_2, which refers to the T_1XP_2, the objective is to add the repair detection at the Rpd and the EdP regions. Table 2 provide an example of labelling reference (here the selected disfluence is a repetition). It underlines the labelling strategy and the elements of a disfluent region to be detected according to the objectives defined within the two detection tasks.

4.2. Features and patterns

We recreated a baseline system based on the one described in [17], using the same features and the same type of model. The patterns are based on the words and applied in a window of -2/+2 words. Linguistic features (Table 3) involve purely lexical features such from part-of-speech features (provided by a French version of the `Brill` POS tagger [27]) to dialogue patterns including speaker turn but also speaker identity. We discretized all the continuous attributes with the tool `discretize4crf` [1]. Table 4 lists the selected parameters classically employed to acoustically characterise speech. Using the LIMSI ASR system [28], an alignment between the audio

1. https://gforge.inria.fr/projects/discretize4crf/.

T_1XP$_1$: I pay by$_{B.Rpd-Edp}$ uh$_{I.Rpd-Edp}$ well$_{I.RpdEdp}$ by card.	**T_1XP$_2$:** I pay by$_{B.Rpd}$ uh$_{B.Edp}$ well$_{I.Edp}$ by card.
T_2XP$_1$: I pay by$_{B.Rpd-Edp}$ uh$_{I.Rpd-Edp}$ well$_{I.Rpd-Edp}$ by$_{B.Rpr}$ card.	**T_2XP$_2$:** I pay by$_{B.Rpd}$ uh$_{B.Edp}$ well$_{I.Edp}$ by$_{B.Rpr}$ card.

TABLE 2 – Examples of disfluency annotations and their differences according to the 4 tasks $\mathbf{T}_i\mathbf{XP}_j$.

signal and the manual transcription at a phonetic level was produced. This alignment was used to extract acoustic parameters.

Lexical sequences (unigrams/bigrams, [-2,+2] window)
- word inflected form ;
- word part-of-speech features.
Regular expressions (unigrams, current word only)
- Prefixes/Suffixes (positions $1:4$).
Yes/No binary patterns (unigrams, [-1,+1] window)
- contains a capital letter ? begins with a capital letter ?
- all in uppercase ?
- contains a punctuation ? is a punctuation ?
- contains a punctuation (except first/last character) ?
- contains a number ? is a number ?
Extra-lexical sequences
- rank of the turn speaker into the conversation ;
- rank of the word into the turn speaker ;
- speaker class ; speaker gender.

TABLE 3 – Patterns for the model based on linguistic features.

Acoustic sequences (unigrams/bigrams, [-2,+2] window)
- word pronunciation (phonemic transcription) ;
- word duration ;
- word number of phonemes ;
- average duration of the word phonemes.
Pitch and formants (unigrams/bigrams, [-2,+2] window)
- F0, F1, F2, F3, F4 mean of the word phonemes ;
- F0, F1, F2, F3, F4 mean of the word vowels ;
- FO $\Delta_{max-min}$ of the word phonemes (raw and normalised by number of syllables) ;
- FO $\Delta_{end-beg}$ of the word vowels (raw and normalised by number of syllables).

TABLE 4 – Patterns for the model based on acoustic features.

5. Experiments and results

We present in this section the results obtained by the four versions of the CRF model within the different experiments corresponding to the two defined tasks. The evaluation is done on the VOXDISS test corpus. The metrics used in our experiments correspond to classic evaluation measures Precision, Recall, F-score and Slot Error Rate or SER [29] [2]. Section 5.1 and Section 5.2 present the results obtained respectively for Task-I and Task-II. Finally, we analyse and discuss them in Section 5.3.

5.1. Task-I : Reparandum and editing phase detection

Table 5 and Table 6 present the results obtained respectively for T_1XP_1 and for T_1XP_2. Generally speaking, one can observe that the acoustic patterns (CRF_A) alone never outperform the linguistic (CRF_L) or even the simpliest ones (BSL). Nevertheless mixed patterns (CRF_LA) offer better performance than the mono-type features ones. For example, one can observe that the F-score of the T_1XP_1 goes from 16.3% with the acoustic patterns to 36.2% with the mixed patterns. The behaviour of the F-score of the T_1XP_2 is similar, going from 18.3% with acoustic patterns to 37.5% with mixed features. Acoustic features lead to very high SER (beyond 83.0% for all tasks and experiments). They are not a good indicator for finding the beginning or ending Rpd/EdP or Rpd-Edp sequences.

Measure	System			
	BSL	**CRF_L**	**CRF_A**	**CRF_LA**
P	**0.564**	0.508	0.280	0.529
R	0.237	0.233	0.115	**0.275**
F	0.333	0.319	0.163	**0.362**
SER	0.769	0.781	0.956	**0.752**

TABLE 5 – Evaluation of T_1XP_1 : Rpd-Edp detection.

Meas.	Label	System			
		BSL	**CRF_L**	**CRF_A**	**CRF_LA**
P	Rpd	0.574	0.592	0.327	**0.600**
	Edp	0.500	**0.550**	0.400	0.500
	all	0.567	0.586	0.336	**0.587**
R	Rpd	0.246	0.276	0.127	**0.291**
	Edp	0.118	**0.216**	0.118	0.196
	all	0.226	0.266	0.125	**0.276**
F	Rpd	0.345	0.377	0.183	**0.392**
	Edp	0.190	**0.310**	0.182	0.282
	all	0.323	0.366	0.183	**0.375**
SER	all	0.817	**0.774**	0.972	0.779

TABLE 6 – Evaluation of T_1XP_2 : Rpd vs. Edp detection.

In Task-I, we compare the two experiments in order to answer the following question : does the isolation of the Edp from the Rpd improve the detection of the disfluent area ? Considering CRF_LA, the best results are obtained with a distinctive identification of Rpd and Edp (T_1XP_2) : Precision is at 58.7% ; the recall of both strategies are quite similar (27.5% vs. 25.6%). Nevertheless the best SER is obtained confusing Rpd and Edp (75.2% in T_1XP_1 and 77.9% in T_1XP_2, both with CRF_LA) : T_1XP_2 contain more insertion and deletion errors as T_1XP_1 presents more substitution errors (9.4% of substitutions for T_1XP_2 vs. 14.5% for T_1XP_1). Corrects are equivalent for both experiments. One can observe these substitutions for T_1XP_1 are border errors only (never type). Considering our main objectives : to improve speech data readability (with an automatic removal of Rpd and Edp), we want to identify disfluent regions as accurately as possible. However, as much as T_1XP_1 presents less numerous both insertion and deletion errors, this result is more interesting : we want to avoid both deleting of well formed segments (linked to insertion errors) and conservation of disfluent segments (suppression errors).

5.2. Task-II : Detection and structuring of the entire edit disfluency area

The results are presented in Table 7 for T_2XP_1 and in Table 8 for T_2XP_2. The use of mixed patterns seems to be the best model for finding edit disfluency sequences : F-score for T_2XP_1 goes from 29.4% with CRF_A to 40.5% with CRF_LA

2. SER measures errors of insertions, substitutions (borders and types) and deletes. This measure is similar to the Word Error Rate, used in ASR.

and for T_2XP_2 goes from 30.3% with CRF_A to 37.4% with CRF_LA. However, the gain obtained for Task-II from the baseline is relative, as n-grams of words offer better Precision : with CRF_LA, it goes from 59.4% to 65.4% for T_2XP_1 and from 67.0% to 68.9% for T_2XP_2. Nevertheless mixed pattern obtain the best Precision for the EdP sequence (80.0%).

Meas.	Label	System			
		BSL	**CRF_L**	**CRF_A**	**CRF_LA**
P	Rpd-Edp	**0.596**	0.570	0.429	0.559
	Rpr	**0.723**	0.681	0.547	0.636
	all	**0.654**	0.622	0.483	0.594
R	Rpd-Edp	0.225	0.233	0.183	**0.252**
	Rpr	0.283	**0.302**	0.245	0.297
	all	0.251	0.264	0.211	**0.297**
F	Rpd-Edp	0.327	0.331	0.257	**0.347**
	Rpr	0.407	**0.418**	0.339	0.405
	all	0.363	0.370	0.294	**0.405**
SER	all	0.790	0.776	0.840	**0.772**

TABLE 7 – Evaluation of T_2XP_1 : detection of the entire area of edit disfluencies, Rpd-Edp as a single sequence vs. Rpr.

Meas.	Label	System			
		BSL	**CRF_L**	**CRF_A**	**CRF_LA**
P	Rpd	**0.656**	0.598	0.477	0.632
	Edp	0.778	0.588	0.538	**0.800**
	Rpr	**0.718**	0.674	0.582	0.694
	all	**0.689**	0.630	0.526	0.670
R	Rpd	0.235	0.228	0.198	**0.250**
	Edp	0.137	0.196	0.137	**0.235**
	Rpr	0.264	**0.283**	0.250	0.278
	all	0.237	0.247	0.213	**0.260**
F	Rpd	0.346	0.330	0.280	**0.358**
	Edp	0.233	0.294	0.219	**0.364**
	Rpr	0.386	**0.399**	0.350	0.397
	all	0.353	0.355	0.303	**0.374**
SER	all	0.792	0.796	0.830	**0.763**

TABLE 8 – Evaluation of T_2XP_2 : detection of the entire area of edit disfluencies, Rpd, Edp as distinct sequences vs. Rpr.

With Task-II, mirroring Task-I, we compare two strategies to answer the question : what is the better strategy to detect an edit disfluency ? inclusion or isolation of the editing phase from the reparandum ? First, analysing results for the **repair**, one can observe that better models are the same for both strategies : best F-score is obtained with the linguistic model. However, the baseline offers better Precision : for T_2XP_1, 72.3% for BSL vs. 62.2% with CRF_L ; for T_2XP_1, BSL presents a 71.8% Precision vs. 67.4% with CRF_L. In a global way, considering Rpd and Edp detection in a single sequence is the best strategy to detect repairs (with an F-score at 41.8% when using linguistic patterns alone). Second, analysing results for the **disfluent area** : trends are similar between T_2XP_1 and T_2XP_2, following the same scheme as for Rpr detection in Precision. But unlike results for Rpr detection, best F-score results are obtained with the combination of linguistic and acoustic features (35.8% for Rpd and 36.4% for Edp in T_2XP_2). Associated with Rpr detection, the best strategy for disfluent phase identification is to consider Rpd and Edp as two distinct sequences (T_2XP_2).

Finally, even if the best F-score results for detecting the disfluent area in Task-II are obtained by distinguishing Rpd and Edp, we can affirm that the best strategy to detect and structuring an edit disfluency is to identify two associated sequences : Rpd and Edp as a single sequence for the disfluent area vs. the associated Rpr (T_2XP_2), using mixed acoustic and linguistic patterns.

5.3. Task-I *vs.* Task-II : Does the repair detection help to detect reparandum and editing phase ?

Considering the best strategy to detect an edit disfluency (Rpd-Edp on the one hand and Rpr on the second hand, with CRF_LA), one can observe that balance between Precision and Recall is reversed : we earn 3 points of Precision but 2.3 points of Recall are lost (from T_1XP_1 to T_2XP_1). This loss also appears for Rpd sequence detection considering the second strategy (distinguish Rpd from Edp). However, Rpr association clearly outperforms Edp detection results : between T_1XP_2 and T_2XP_2, Precision goes from 50% to 80% and Recall from 19.6% to 23.5%. Repair detection definitively helps editing phase detection.

6. Conclusion and future work

We have provided a model for the automatic detection of edit disfluencies in a call-centre corpus in French based on Conditional Random Fields. Two related tasks were considered according to the dedicated application challenges : the first task is dedicated to improve speech analytics interfaces and the second task consists in a challenging issue for call-centre data mining applications.

We showed that depending on the considered evaluation measure (Precision, Recall, F-score and Slot Error rate), the best emerging strategies may be different. In terms of F-score, the use of both linguistic and acoustic features in the defined disfluency patterns allowed us to obtain the best results for both tasks and experiments. Acoustic cues give the poorest outcomes, which is expected since we work on manual transcriptions. Acoustic cues give the poorest outcomes, which is expected since we work on manual transcriptions. We plan to apply our method to automatic transcriptions, especially to assess the impact of speech recognition errors in the disfluency detection system. Best results are obtained with the identification of the entire area of edit disfluencies using both linguistic and acoustic features, considering the reparandum and the editing phase as a single sequence vs. the associated repair (with a global Precision of 59.4%). The results are quite promising considering the size of the available annotated data and the richness of call-centre data in terms of disfluent phenomena (from repetitions to complex disfluencies).

Future work will be dedicated, first, to an in-depth analysis of the system outputs according to the considered type of disfluencies and, second, to a study of the patterns used by the CRF that are the most relevant for the decision process. Besides, we plan to refine our CRF-based model by distinguishing the different classes of edit disfluencies (restarts, complex disfluencies, etc.). Even though they are not as good as those reported in the literature, the obtained results are quite promising considering that the proposed method tackles a more complex task in a different language. Moreover, it will be crucial to define protocols for the comparison of our approach to other approaches, which should allows us to draw strongest conclusions.

7. Acknowledgements

This work was partially funded by the CIFRE convention 2011/0916.

8. References

[1] E. E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, Berkeley University of California, 1994.

[2] H. H. Clark and J. E. Fox Tree, "Using uh and um in Spontaneous Speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[3] C. Clavel, G. Adda, F. Cailliau, M. Garnier-Rizet, A. Cavet, G. Chapuis, S. Courcinous, C. Danesi, A.-L. Daquo, M. Deldossi, S. Guillemin-Lanne, M. Seizou, and P. Suignard, "Spontaneous Speech and Opinion Detection : Mining Call-centre Transcripts," *Language Resources & Evaluation*, vol. 1, p. 40, 2013.

[4] F. Cailliau, A. Giraudel *et al.*, "Enhanced Search and Navigation on Conversational Speech," *Proceedings of SSCS*, pp. 66–70, 2008.

[5] R. J. Lickley, "Detecting Disfluency in Spontaneous Speech," Ph.D. dissertation, University of Edinburgh, 1994.

[6] S. Strassel, *Simple Metadata Annotation Specification*, Linguistic Data Consortium, 2004.

[7] W. J. M. Levelt, "Monitoring and Self-repair in Speech," *Cognition*, vol. 14, pp. 41–104, 1983.

[8] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based Technique for Automatic Filled-pauses Detection in Spontaneous Spoken English," in *Proceedings of ICASSP*, 2009.

[9] M. Kaushik, M. Trinkle, and A. Hashemi-Sakhtsari, "Automatic Detection and Removal of Disfluencies from Spontaneous Speech," in *Proceedings of the 13^{th} Australasian International Conference on Speech Science and Technology*, 2010, pp. 98–101.

[10] M. Snover, B. Dorr, and R. Schwartz, "A Lexically-Driven Algorithm for Disfluency Detection," in *Proceedings of HLT-NAACL*, 2004.

[11] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting Structural Metadata with Decision Trees and Transformation-Based Learning," in *Proceedings HLT/NAACL*, 2004, pp. 137–144.

[12] Y. Liu, E. E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1526 – 1540, 2006.

[13] M. Johnson and E. Charniak, "A TAG-based Noisy Channel Model of Speech Repairs," in *Proceedings of ACL*, 2004.

[14] S. Zwarts and M. Johnson, "The Impact of Language Models and Loss Functions on Repair Disfluency Detection," in *Proceedings of ACL*, 2011, pp. 703–711.

[15] K. Georgila, "Using Integer Linear Programming for Detecting Speech Disfluencies," in *Proceedings of HLT/NAACL*, 2009, pp. 109–112.

[16] M. S. Rasooli and J. Tetreault, "Joint Parsing and Disfluency Detection in Linear Time," in *Proceedings of EMNLP*, 2013, pp. 124–129.

[17] X. Qian and Y. Liu, "Disfluency Detection Using Multi-step Stacked Learning," in *Proceedings of HLT/NAACL*, 2013, pp. 820–825.

[18] K. Georgila, N. Wang, and J. Gratch, "Cross-Domain Speech Disfluency Detection," in *Proceedings of SIGDial*, 2010.

[19] M. Constant and A. Dister, *Spoken Communication*. Cambridge Scholars Publishing, 2010, ch. Automatic Detection of Disfluencies in Speech Transcriptions, pp. 259–272.

[20] J.-L. Bouraoui and N. Vigouroux, "Traitement automatique de disfluences dans un corpus linguistiquement contraint," in *Actes de TALN*, 2009.

[21] K. Peshkov, L. Prvot, S. Rauzy, and B. Pallaud, "Categorizing Syntactic Chunks for Marking Disfluent Speech in French Language," in *Proceedings of DiSS*, 2013, pp. 59–62.

[22] M. Garnier-Rizet, G. Adda, F. Cailliau, J.-L. Gauvain, S. Guillemin-Lanne, and L. Lamel, "CallSurf : Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content," in *Proceedings of LREC*, 2008, pp. 2623–2628.

[23] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech," in *Proceedings of LREC*, 1998, pp. 1373–1376.

[24] J. D. Lafferty, A. McCallum, and F. C. N. Pereira., "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of International Conference on Machine Learning*, 2001, pp. 282–289.

[25] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," in *Proceedings ACL*, 2010, pp. 504–513, Wapiti Web site : http://wapiti.limsi.fr/.

[26] M. Dinarelli and S. Rosset, "Models Cascade for Tree-Structured Named Entity Detection," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011, pp. 1269–1278.

[27] A. Allauzen and H. Bonneau-Maynard, "Training and Evaluation of POS Taggers on the French MULTITAG Corpus," in *Proceedings of LREC*, 2008.

[28] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," Hong Kong, April 2003, pp. I–212–215. [Online]. Available : ftp://tlp.limsi.fr/public/ica03cts.pdf

[29] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance Measures For Information Extraction," in *Proceedings of DARPA Broadcast News Workshop*, 1999, pp. 249–252.