SPONTANEOUS SPEECH CHARACTERIZATION AND DETECTION IN LARGE AUDIO DATABASE

R. Dufour, V. Jousse, Y. Estève

LIUM - Université du Maine
Le Mans, France

F. Béchet, G. Linarès

LIA - Université d'Avignon
Avignon, France

ABSTRACT

Processing spontaneous speech is one of the many challenges that Automatic Speech Recognition (ASR) systems have to deal with. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start) and many studies have focused on the detection and the correction of these disfluencies. In this study we define *spontaneous speech* as *unprepared speech*, in opposition to *prepared speech* where utterances contain well-formed sentences close to those that can be found in written documents. Disfluencies are of course very good indicators of *unprepared speech*, however they are not the only ones: ungrammaticality and language register are also important as well as prosodic patterns. This paper proposes a set of acoustic and linguistic features that can be used for characterizing and detecting spontaneous speech segments from large audio databases. To better define this notion of unprepared speech, a set of speech segments representing an 11 hour corpus (French Broacast News) has been manually labelled according to a level of spontaneity. We present an evaluation of our features on this corpus, describe the correlation between the Word-Error-Rate obtained by a state-of-the-art ASR decoder on this BN corpus and the level of spontaneity and finally propose a strategy that takes advantage of this information in order to improve the ASR performance.

Index Terms— spontaneous speech characterization, spontaneous speech detection, automatic speech recognition

1. INTRODUCTION

Information Extraction (IE) from large audio databases requires to extract the structure of audio documents as well as their linguistic content. One part of this structuration process is to add punctuations and sentence boundaries to the automatic transcriptions of the speech segments detected. This segmentation process is very important for many tasks like speech summarization, speech-to-speech translation or the *distillation* task as defined in the GALE program [1]. Adding this structure to the automatic transcripts is a very challenging task when processing spontaneous speech as this kind of speech is characterized by ungrammaticality and disfluencies. It is therefore useful to detect spontaneous speech segments at an early stage in order to adapt the ASR and structuration processes to this particular kind of speech. This is the goal of this study.

Spontaneous speech occurs in Broadcast News (BN) data under several forms: interviews, debates, dialogues, etc. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start) and many studies have focused on the detection and the correction of these disfluencies [2, 3] as pointed out by the recent NIST Rich Transcription Fall 2004 blind evaluation. All these studies show an important drop in performance between the results obtained on reference transcritptions and those obtained on automatic transcripts. This can be explained by the noise generated by ASR systems on spontaneous speech segments with higher Word Error Rate (WER) values than on *prepared* speech. Indeed high WER values are obtained by state-of-the-art ASR systems when transcribing data likely to contain a lot of spontaneous speech like conversational speech or meeting recordings. One goal of this study is to closely illustrate this link between WER and spontaneous speech.

In addition to disfluencies, spontaneous speech is also characterized by ungrammaticality and a language register different from the one that can be found in written texts [4]. Depending on the speaker, the emotional state and the context, the language used can be very different. In this study we define *spontaneous speech* as *unprepared speech*, in opposition to *prepared speech* where utterances contain well-formed sentences close to those that can be found in written documents. We propose a set of acoustic and linguistic features for characterizing *unprepared speech*. The relevance of these features is estimated on an 11 hour corpus (French Broadcast News) manually labelled according to a level of spontaneity in a scale from 1 (clean, prepared speech) to 10 (highly disfluent speech, almost not understoodable). We present an evaluation of our features on this corpus, describe the correlation between the Word-Error-Rate obtained by a state-of-the-art ASR decoder on this BN corpus and the level of spontane-

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-06-MDCA-006.

ity and finally propose a strategy that takes advantage of this information in order to improve the ASR performance.

2. SPONTANEOUS SPEECH CHARACTERIZATION

2.1. Levels of spontaneity

By defining spontaneous speech as *unprepared speech*, we follow a definition proposed by [5] that defined a spontaneous utterance as: "a statement conceived and perceived during its utterance". This definition illustrates the subjectivity of the classification prepared/spontaenous speech. Ideally, to annotate a speech corpus with labels representing the spontaneity of each speech segment, we would have to ask each speaker to annotate his own utterances. This is of course not feasible, however we followed this definition by defining an annotation protocol based on the perception by a human judge of a *level of spontaneity* for a given speech segment. Our approach was to manually tag a corpus of speech segments with a set of ten labels corresponding each to a spontaneity level: grade 1 stands for prepared speech, almost similar to read speech, and grade 10 stands for very disfluent speech, almost not understandable. This approach allows us to subjectively choose where the limit between spontaneous and prepared speech is placed. In the experiment we considered 3 classes: *prepared speech* corresponding to grade 1; *low spontaneity* corresponding to the grades 2 to 4; and *high spontaneity* corresponding to the grade 5 and over.

Two human judges have annotated a speech corpus by listening to the audio recordings. The corpus was cut into segments thanks to an state-of-art automatic segmentation and diarization process [6]. No transcriptions were provided to the annotators. In order to evaluate inter-annotator agreement for this specific tagging task on the 3 classes presented above, we computed the Kappa coefficient of agreement [7] on one hour of Broadcast News. The coefficient obtained was very high: 0.852 — a value greater than 0.8 is usually considered as excellent [8].

Then, they have annotated the remaining corpus separately. One of the problems encountered was that spontaneous speech segments can occur everywhere, not only in conversational speech, in the middle of very *clean* utterances. Similarly even conversational speech can contain segments that can be considered as prepared speech. To take this into account, we decided to evaluate each segment independently: a spontaneous segment can be surrounded by many prepared ones.

The corpus obtained after this labelling process is made of 11 files containing French Broadcast News data from 5 different media (France Culture, France Inter, France Info, Radio Classique, RFI). The files were chosen for being likely to contain spontaneous speech according to the kind of radio show broadcast. The total duration is 11h37 for a total of 3322 segments (after removal of the non speech segments: music, jingles, ...). Among these segments, 1142 were annotated with the *prepared speech* label, 1175 with the *low spontaneity* label and 1005 with the *high spontaneity* label.

2.2. Acoustic and linguistic features

In parallel to the subjective annotation of the corpus presented in the previous section, we introduce now the features used to describe speech segments, relevant to characterize the spontaneity of them, and on which an automatic classification process can be trained on our annotated corpus. This problem has been studied recently as a specific task from the Rich Transcription Fall 2004 blind evaluation which was focused on the detection of speech disfluencies. Some approaches use only linguistic features [3], both linguistic and prosodic features [9], or linguistic and more general acoustic features [10].

In this paper we use two sets of features: acoustic features related to prosody and linguistic features related to the lexical and syntactic content of the segments. We combine both of them in order to characterize the spontaneity level of a speech segment: this task is different from the speech disfluency detection task as spontaneous speech segments don't necessarily contain disfluencies. For example, they can also be characterized by a high variation in the speech rate. The features used in this study are briefly presented in the next section.

2.2.1. *Prosodic features*

The prosodic features used are related to vowel duration and phonetic rate, as presented below.

Duration: following previous work describing the link between prosody and spontaneous speech [11], we use two features: vowel duration and the lengthening of a syllable at the end of a word. This last features has been proposed in [12] and is associated to the concept of *melism*. In addition to the average durations, their variance and standard deviation are also added as features in order to measure the dispersion of the durations around the average.

Phonetic rate: previous studies [12] have shown the correlation between the variations of speech rate and the emotional state of a speaker. Following this idea we use as feature an estimate of the speech rate, by word or by speech segment, in order to observe its impact on the spontaneity of the speech. We estimate the phonetic rate in two ways: the variance of the phonetic rate for each word, and the average of the phonetic rate on the whole segment, including pauses and fillers.

2.2.2. *Linguistic features*

The main characteristic of spontaneous speech is the concept of *speech disfluencies*. They can be categorized as filled pause, repetition, repair and false start. A lot of studies have been focused on their description at the acoustic [11] or lexical level [13]. We use two features representing them in the description of the speech segments:

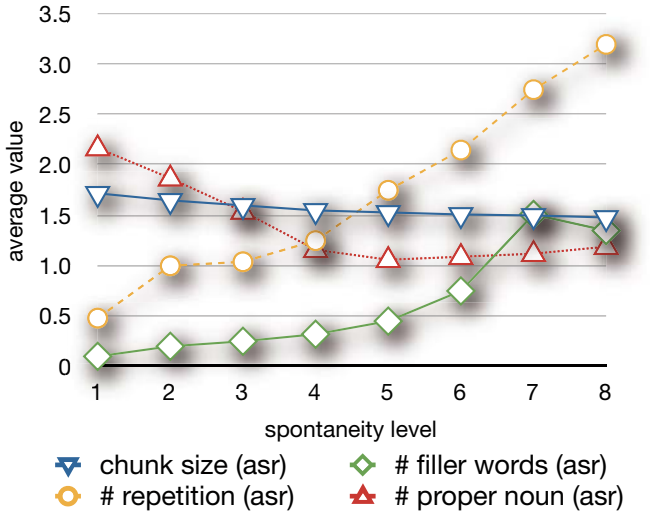

Fig. 1. Linguistic feature average values according to the degree of spontaneity on the manually labeled corpus

- filled pause: the ASR lexicon contains several symbols, filler words, for representing filled pause in French, like *euh*, *ben* or *hum*. The number of occurences of all of them in a segment is the first feature.

- repetition and false start: we use here a very simple feature counting the number of 1-gram and 2-gram repetition in a segment.

As shown by [4] on BN data, spontaneous speech is also characterized at the linguistic level by other phenomenon than filled pause or repetition. Agrammaticality and language register are also very characteristic of unprepared speech. In order to capture this link between spontaneity on one side and lexicon and syntax on the other side, we apply to the transcriptions of audio segments a shallow parsing process including a POS tagging and a syntactic chunking process and use the following features to describe them:

- bags of n-grams (from 1 to 3-grams) on words, POS tags and syntactic chunk categories (noun phrase, prepositional group;

- average length of syntactic chunks on the segment.

Moreover, as presented in [14], a high number of occurrences of proper nouns in a speech segment can be informative to characterize prepared speech: this information is used in this work.

Figure 1 shows the correlation between the level of spontaneity assigned to the speech segment of our corpus and the linguistic features presented. Although these numbers are obtained on automatic transcripts with a high WER on the most spontaneous segments, there is a clear increase for both disfluency features between clean and spontaneous speech. Although the variation of the average chunk size is limited, figure 1 shows a reduction of this size, from 1.7 down to 1.4 words on average per chunk between clean speech and very spontaneous one.

3. AUTOMATIC DETECTION OF SPONTANEOUS SPEECH SEGMENTS

The features presented in the previous section are evaluated on our labeled corpus with a classification task: labeling speech segments according to the three spontaneity levels: *prepared speech*, *low spontaneity* or *high spontaneity* label. The classification tool used is *BoosTexter* based on the AdaBoost algorithm [15]. This is a large-margin classifier based on a boosting method of *weak* classifiers. The weak classifiers are given as input. They can be the occurrence or the absence of a specific word or n-gram (for the linguistic features) or a numerical value (for the acoustic features, the disfluencies and the average chunk size). At the end of the training process, the list of the selected classifiers is obtained as well as the weight of each of them in the calculation of the classification score for each speech segment to process.

The corpus (as described in 2.1) is made of 11 audio files. For the experiments, we used the *Leave One Out* method: 10 files used for training, 1 for the evaluation and this process is repeated until all files have been evaluated.

Five conditions were evaluated:

- Linguistic features only on reference transcription *ling(ref)*

- Linguistic features only on automatic transcription *ling(asr)*

- Acoustic features only on automatic transcription *acou(asr)*

- All features on automatic transcription *all(asr)*

- Consensus between *ling(asr)*, *acou(asr)* and *all(asr)*: *con(asr)*

Table 1 presents the detection results (precision and recall) for each spontaneity label. As we can see the detection performance on the *low spontaneity* segments is low, this is not suprising as these segments can be easily misclassified as *prepared speech* one one side or *high spontaneity* on the other side. The recall measures are correct however the precision measures remain quite low, even for the *high spontaneity* label. As we can see the drop between the performance achieved on the reference transcriptions using linguistic features and the automatic transcriptions, due to ASR errors, is compensated by the acoustic features that are more robust to ASR errors: the use of a classifier based on all the acoustic

prepared speech					
Features	ling(ref)	ling(asr)	acou(asr)	all(asr)	con(asr)
Precision	51.4	47.1	54.1	54.7	57.6
Recall	77.0	72.9	67.3	66.4	69.8
low spontaneous					
Features	ling(ref)	ling(asr)	acou(asr)	all(asr)	con(asr)
Precision	37.1	33.4	39.5	41	42.3
Recall	70.1	17.6	35.5	34.8	36.5
high spontaneous					
Features	ling(ref)	ling(asr)	acou(asr)	all(asr)	con(asr)
Precision	53.0	59.8	61.6	62.1	**66.1**
Recall	75.0	55.3	50.8	57.4	**58.9**

Table 1. Precision, recall and F-measure in the classification of the speech segments according to 3 categories: *prepared speech* (grade 1), *low spontaneity* (grade 2-4) and *high spontaneity* (grade 5-10)

and linguistic features extracted automatically (*all(asr)*) improves performances, especially for precision. More, a consensual decision of the 3 classifiers based on features automatically extracted allows to achieve a better 66.1% precision in the high spontaneous speech detection with a 58.9% recall measure.

4. USING SPONTANEOUS SPEECH DETECTION FOR IMPROVING ASR

The main application of this study is to use the information about the level of spontaneity of a speech segment in order to improve ASR performance. For example word pronunciation is affected in a spontaneous speech context, and previous studies [16] have focused on improving word pronunciation for spontaneous speech recognition. In this paper, we aim to demonstrate the benefits of spontaneous speech characterization and detection in audio database in order to improve ASR performance. To this purpose, we propose to investigate an approach based on Language Model (LM) adaptation.

4.1. ASR description

Experiments on speech recognition were made by using the LIUM ASR system based on the CMU Sphinx 3.x decoder [6]. It is a three-passes system: a first pass uses a trigram language model and generic acoustic models (one for each of the four gender/band conditions: female/male + studio/telephone), a second pass uses the best hypothesis of the first pass to adapt the acoustic models using SAT and CMLLR, and the last pass consists in rescoring with a quadrigram language model a word-graph generated during the second pass. This system was ranked second in the French ASR evaluation program ESTER [17] on broadcast news recordings.

4.2. Generic LM description

The training data provided during the ESTER evaluation campaign is used to train our generic LM. It is made of manual transcriptions (80h) of broadcast news data, but the main part of the data comes from the French newspaper *Le Monde*. The amount of very spontaneous speech is limited in this corpus, therefore the main contribution to this LM comes from written texts.

Using the vocabulary (65K words) built for the LIUM participation to the ESTER program, the two data sets (*Le Monde* + BN manual transcriptions) are used to train the baseline trigram and quadrigram LMs. To estimate and interpolate these LMs, the SRILM toolkit [18] is used. Each language model is a backoff model, using the Kneser-Ney modified discounting method with interpolation for low-order *n-grams*.

4.3. Adapted LM estimation

Because the BN corpus presented in section 2.1 contains a lot of spontaneous speech, there is a mismatch between this genereic LM and our spontaneous speech corpus. Therefore we have tested two approaches based on LM adaptation for dealing with this mismatch:

1. the first approach consists in selecting, from the training corpus of the generic LM, the sentences containing transcriptions of spontaneous speech, then boosting the probabilities of the corresponding *n-grams* in the generic LM;

2. in the second approach we use an external corpus, which has no link with broadcast news data, but which contains only transcriptions of very spontaneous speech with the explicit annotation of speech disfluencies. Here, the goal is to adapt the generic LM by integrating new information, relevant to spontaneous speech, provided by another knowledge source.

For the two approaches, the LM adaptation consists in interpolating linearly the initial generic LM with the new LM trained on the selected corpus.

For the first approach we have automatically extracted transcriptions of spontaneous speech from the training corpus of the baseline LM. This extraction was made using the spontaneous speech detection process based on linguistic features presented in section 2.2. This sub-corpus of the baseline LM was used to estimate the spontaneous speech LM, called *sponta(base)*.

For the second approach we used the manual transcriptions of open conversations made available through the PFC (*Phonologie du Francais Contemporain*) project [1]. The PFC project [19] involves over thirty researchers from a variety of countries and aims at the recording, partial transcription and analysis of over 500 speakers from the francophone

[1] http://www.projet-pfc.net

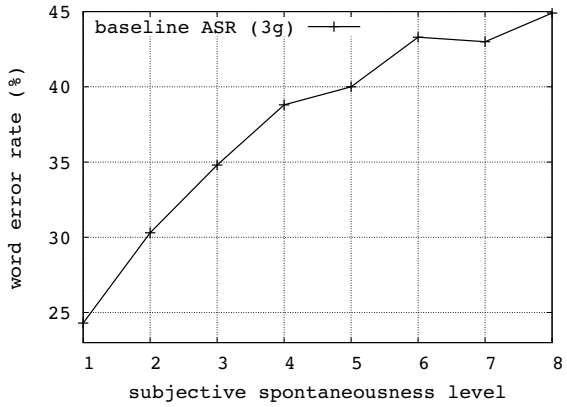

Fig. 2. Word error rate according to the subjective spontaneous levels

world on the basis of a common protocol. The audio recordings are made mostly of open conversations. This corpus of manual transcriptions of conversations constitutes an interesting knowledge source for modeling spontaneous speech phenomenons: we have extracted from it 26K sentences with 285K word occurrences in order to build our spontaneous speech LM.

4.4. Experiments

The training data of the ASR system is not included in the 11 audio files described in section 2.1. Therefore in order to evaluate the correlation between WER and the subjective levels of spontaneity described in section 2.1, we have transcribed these 11 files with the baseline ASR system. Figure 2 shows that there is a real correlation between WER and subjective levels of spontaneity: as soon as a speech segment is not perceived by the human annotator as perfectly uttered (*i.e.* for a level greater than 1), the WER increases. For spontaneity levels 6, 7 and 8, no significant difference was observed in terms of WER, but the WER is yet very high. Notice that there was no speech segment annotated with a spontaneity level greater than 8 in this corpus.

For the following experiments with the ASR system, the 11 audio files are divided into two corpora: a development corpus (7 files) and a test corpus (4 files, about 4h15 duration). As presented in section 2.1 the different spontaneity levels are grouped into 3 classes: *prepared speech* (level 1), *low spontaneous speech* (levels 2-3-4), and *high spontaneous speech* (levels ≥ 5). Table 2 shows the distribution of these segments in the test corpus.

The development corpus is used to optimize the interpolation coefficient between the generic LM and the spontaneous

	prepa.	low sponta.	high sponta.	global
# segments	1314	1152	1715	4181
# words	13493	12218	19292	45008

Table 2. Number of speech segments and words in the text corpus according two the three classes of spontaneity.

speech LMs. Two adapted LM are tested:

- LM *base+sp.(base)* which is the result of the interpolation between the generic LM and the LM called *sponta(base)*;
- LM *base+pfc* which is the result of the interpolation between the generic LM and the LM estimated on the PFC corpus.

LM	prepared	low sp.	high sp.
baseline	156	193	203
base+sp.(base)	186	199	180 (-11.33%)
base+pfc	171	184	164 (-19.21%)

Table 3. Perplexity according to the LM and the spontaneity classes

Table 3 shows the results in terms of perplexity on the test corpus according to the spontaneity class. As we can see using a LM specific to spontaneous speech reduces significantly the perplexity for the classes *low spontaneity* and *high spontaneity* while as expected the perplexity increases for the class *prepared speech*. It is interesting to notice that the PFC corpus brings a gain only on the *high spontaneity* class. This can be explained by the thematic mismatch between this corpus and the broadcast news data: only highly disfluent speech can benefit from this corpus despite this mismatch.

LM	prepared	low sp.	high sp.	global
baseline	21.4	31.3	41.2	32.6
base+sp.(base)	22.3	31.7	40.5	32.6
base+pfc	21.7	31.8	40.1	32.4

Table 4. Word error rate according to the LM and the spontaneity classes

WER results are presented in table 3. These WER are relatively high because the files chosen for the experiments contain much more spontaneous speech than the ones used during the ESTER evaluation program. Similar to the perplexity results, we can notice that using the two adapted LMs reduces WER on the highly spontaneous speech while they increase the WER on other kinds of speech. The best LM on

spontaneous speech is the LM estimated on the PFC data containing a major part of conversational speech. But it is interesting to notice that a decrease of WER can also be obtained without adding new data, by 'boosting' specific data automatically selected for containing spontaneous speech transcriptions and already included in the initial training corpus used for the generic LM.

The improvement in WER is not as significant as the one obtained with the perplexity measure: other phenomena than language modeling have to be taken into account and modeled for the ASR processing, as the word pronunciation, the variability of the phone duration, and other typical features as the ones seen in section 2.2. These are the immediate perspectives of this work.

5. CONCLUSION

We propose a set of acoustic and linguistic features that can be used for characterizing and detecting spontaneous speech segments from large audio databases. To better define this notion of unprepared speech, a set of speech segments representing an 11 hour corpus (French Broacast News) has been manually labelled according to a level of spontaneity: the correlation between the Word-Error-Rate and the level of spontaneity obtained by LIUM state-of-the-art ASR decoder on this BN corpus is presented.

The acoustic and linguistic features are evaluated for characterising and detecting spontaneous speech segments: the combination of acoustic and linguistic features extracted from ASR outputs obtains a better precision than the linguistic features extracted from the reference transcriptions alone, for a lower recall measure. But, using consensus between the classifiers based on features automatically extracted, an better 66.1% precision in the detection of high spontaneous speech can be achieved with a 58.9% recall measure. Of course, these results have to be improved: this task is hard, and we do not know a communication presenting better results on this task.

We present also some preliminary experiments which show that this spontaneous speech characterization can be useful in order to adapt ASR model to spontaneous speech effects. We demonstrate that our spontaneous speech detection process can be used in order to select a corpus on which a Language Model dedicated to spontaneous speech can be trained. This method achieved interesting results in term of perplexity and WER reduction on our test corpus.

6. REFERENCES

[1] D. Hakkani-Tur and G. Tur, "Statistical Sentence Extraction for Information Distillation," *ICASSP 2007*, vol. 4, 2007.

[2] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection," *InterSpeech 2005*, 2005.

[3] M. Lease, Johnson M., and E. Charniak, "Recognizing Disfluencies in Conversational Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1566–1573, 2006.

[4] P.B. de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek, "A quantitative study of disfluencies in French broadcast interviews," *Proceeding of the workshop Disfluency In Spontaneous Speech (DISS), Aix-en-Provence, France*, 2005.

[5] D. Luzzati, "Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané," in *MIDL*, Paris, France, 2004.

[6] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based System for French Broadcast News," in *Interspeech 2005*, Lisbon, Portugal, September 2005.

[7] Cohen J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[8] Barbara Di Eugenio and Michael Glass, "The Kappa statistic: A second look," *Computational Linguistics*, vol. 30, no. 1, pp. 95–101, 2004.

[9] Y. Liu, 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, no. 5, pp. 1526–1540, 2006.

[10] J.-F. Yeh and C.-H. Wu, "Edit Disfluencies Detection and Correction Using a Cleanup Language Model and an Alignment Model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1574–1583, 2006.

[11] E. Shriberg, "Phonetic consequences of speech disfluency," *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, pp. 619–622, 1999.

[12] G. Caelen-Haumont, "Perlocutory Values and Functions of Melisms in Spontaneous Dialogue," *Proceedings of the 1st International Conference on Speech Prosody, SP*, pp. 195–198, 2002.

[13] M.H. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," *ICSLP 1996*, vol. 1, 1996.

[14] T. Bazillon, Y. Estve, and D. Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech," in *LREC 2008*, Marrakech, Morroco, 2008.

[15] Robert E. Schapire and Yoram Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[16] C.-K. Lin and L.-S. Lee, "Pronunciation modeling for spontaneous speech recognition using latent pronunciation analysis (LPA) and prior knowledge," in *ICASSP 2007*, Honolulu, Hawaii, USA, 2007.

[17] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Interspeech 2005*, Lisbon, Portugal, September 2005.

[18] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *ICSLP 2002*, Denver, Colorado, USA, 2002, vol. 2, pp. 901–904.

[19] J. Durand, B. Laks, and C. Lyche, "Un corpus numérisé pour la phonologie du français," in *Colloque : La linguistique de corpus*, G. Williams, Ed., Lorient, France, 2005, pp. 205–217.