Local and global models for spontaneous speech segment detection and characterization

Richard Dufour 1, Yannick Estève 1, Paul Deléglise 1, Frédéric Béchet 2

1 LIUM - University of Le Mans, France
firstname.lastname@univ-lemans.fr

2 LIA - University of Avignon, France
firstname.lastname@univ-avignon.fr

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 study1 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. More, we introduce a strategy that takes advantage of a global classification process using a probabilistic model which significantly improves the spontaneous speech detection.

I. INTRODUCTION

Information Extraction (IE) from large audio databases requires to extract the structure of audio documents as well as their linguistic content. 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. Moreover, in order to cluster some documents according to their contents or structure, the presence of spontaneous speech segments should be an interesting descriptor. It is therefore useful to detect spontaneous speech segments at an early stage in order to adapt the ASR, as presented in [1] and structuring 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 transcriptions 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.

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 have already proposed a set of acoustic and linguistic features for characterizing unprepared speech in a previous work [1].

In this new paper, we present some evolutions of this study. Particularly, we present a global approach using global probabilistic modeling which allows to very significantly improve the spontaneous detection performance.

II. SPONTANEOUS SPEECH CHARACTERIZATION

A. Levels of spontaneity

The classification of prepared/spontaneous speech is very subjective. 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. This protocol is based on the perception given by a human judge thanks to a level of spontaneity for a given speech segment. Our approach was to manually tag a corpus of speech segments with a set of eight labels, each one corresponding to a spontaneity level: grade 1 stands for prepared speech, almost similar to read speech, and grade 8 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.

In this paper, we focus particularly on the detection of the high spontaneity class of speech.
Two human judges have annotated a speech corpus by listening to the audio recordings. The corpus was cut into segments thanks to a state-of-art automatic segmentation and diarization process [5]. 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 [6] 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 [7].

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 11821 segments (after removal of the non speech segments: music, jingles, …). Among these segments, 3670 were annotated with the prepared speech label, 4107 with the low spontaneity label and 4044 with the high spontaneity label.

B. Acoustic and linguistic features

In our work, as described in [1], we use three sets of features: acoustic features related to prosody, linguistic features related to the lexical and syntactic content of the segments, and confidence measures. We combine them in order to characterize the spontaneity class of a speech segment: this task is different from the speech disfluency detection task as spontaneous speech segments do not 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.

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 [8], we use two features: vowel duration and the lengthening of a syllable at the end of a word. The latter has been proposed in [9] 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 [9] 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 speech segment, in order to observe its impact on the spontaneity of the speech. We estimate the phonetic rate in two ways: the average and the variance of the phonetic rate on the whole segment, firstly including pauses and fillers, and secondly removing them.

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 [8] or lexical level [10]. We use two features representing them in the description of the speech segments:

- filled pause: the ASR lexicon contains several symbols, filler words, for representing filled pause in French, like euh, ben or hum. The number of occurrences 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 repetitions in a segment.

As shown by [4] on BN data, spontaneous speech is also characterized at the linguistic level by other phenomena than filled pause or repetition. Agrammaticity 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, words and POS tags count.

Moreover, as presented in [11], 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.

3) Confidence measures: Confidence measures are computed scores that expressed reliability of recognition decisions made by ASR system. These scores could be used to characterize spontaneity of speech segments. Indeed, as seen in [1], automatic speech recognition systems have more difficulties to well recognize spontaneous speech segments than prepared speech segments.

III. AUTOMATIC DETECTION OF SPONTANEOUS SPEECH SEGMENTS

To automatically extract the acoustic and linguistic descriptors to categorize speech segments according to class of spontaneity, we used the LIUM ASR system. This ASR system, described in [5], was developed to participate to the French ESTER 2 evaluation campaign on Broadcast News automatic transcription systems.

A. Classification

The features presented in the previous section are evaluated on our labeled corpus with a classification task: labeling speech segments according to the three classes of spontaneity: prepared speech, low spontaneity or high spontaneity label. The classification tool used is icsiboost, an open source tool based on the AdaBoost algorithm like the Boostexter software [12]. This is a large-margin classifier based on a boosting method of weak classifiers.

This classification process, taking into consideration the acoustic and linguistic descriptors presented in section II-B, plus some other descriptors as the duration of speech segments and the number of recognized words proposes a categorization of the speech segments according the three class of spontaneity. Each segment is processed individually.
IV. PROBABILISTIC MODEL FOR GLOBAL DECISION

Our previous approach, presented in [1], takes only into consideration the descriptors which are extracted from within the targeted segment, without taking into consideration information about surrounding segments. In order to improve our approach, we propose to take into account the nature of the contiguous neighboring speech segments. It implies that the categorization of each speech segment from an audio file has an impact on the categorization of the other segments: the decision process becomes a global process. We have chosen to use a classical statistical approach by using a maximum likelihood method.

Let be $s_i$ a tag of the segment $i$, with $s_i \in \{"high spontaneity","low spontaneity","prepared"\}$. We define $P(s_i|s_{i-1}, s_{i+1})$ as the probability of observing a segment $i$ associated to the tag $s_i$ when the previous segment is associated to the tag $s_{i-1}$ and the next segment is associated to the tag $s_{i+1}$. Let be $c(s_i)$ the confidence measure given by the AdaBoost classifier on choosing the tag $s_i$ for the speech segment $i$ according to the values of the descriptors extracted from this segment. $S$ is a sequence of tags $s_i$ associated to the sequence of all the speech segments $i$ (only one tag by segment). The global decision process consists in choosing the tag-sequence hypothesis $\tilde{S}$ which maximizes the global score obtained by combining $c(s_i)$ and $P(s_i|s_{i-1}, s_{i+1})$ for each speech segment $i$ detected on the audio file. The sequence $\tilde{S}$ is computed by using the following formula:

$$\tilde{S} = \arg \max_S c(s_0) \times c(s_n) \times \prod_{i=2}^{n-1} c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$$

where $n$ is the number of speech segments automatically detected in the recording file.

V. EXPERIMENT

The experimental corpus (as described in II-A) is made of 11 audio files from radiophonic recording. 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.

A. ASR performances

The acoustic and linguistic features used as descriptors to characterize the spontaneous speech are issued from the LIUM ASR system. Table I presents the results in terms of word error rate (WER) and normalized cross entropy (NCE) of this ASR system on the experimental data. These data were not included in the training or development corpus of the models used in the ASR system. The WER is the classical metric to evaluate ASR systems, while the NCE is usually used to evaluate the confidence measures provided by an ASR system.

Table I shows that the global performances of the ASR, with a WER of 15% and a NCE of 0.331 are very good for French Broadcast News processing. As it was expected, more the speech is fluent, more the WER is low: from 10.1% for speech segments manually annotated as "prepared" until 28.5% for "high spontaneous" speech segments. Notice the correlation between subjective annotation on spontaneous level and the WER obtained by an ASR system.

B. Automatic categorization and detection of spontaneous speech

In order to measure the information provided by the different kinds of descriptors and the gain provided by the use of a probabilistic model for global decision, 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), use of a probabilistic global model all(asr) results: all + global(asr).

Table II presents the detection results (in terms of precision and recall) for each spontaneity class. As we can see the detection performance on the low spontaneity segments is low, this is not surprising as these segments can be easily misclassified as prepared speech one one side or high spontaneity on the other side.

<table>
<thead>
<tr>
<th>prepared speech</th>
<th>low spontaneous</th>
<th>high spontaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feat.</td>
<td>ling(ref)</td>
<td>ling(asr)</td>
</tr>
<tr>
<td>Prec.</td>
<td>66.1</td>
<td>61.8</td>
</tr>
<tr>
<td>Recall</td>
<td>56</td>
<td>53.0</td>
</tr>
<tr>
<td></td>
<td>acou(asr)</td>
<td>all(asr)</td>
</tr>
<tr>
<td>Prec.</td>
<td>43.8</td>
<td>40.7</td>
</tr>
<tr>
<td>Recall</td>
<td>37.7</td>
<td>31.7</td>
</tr>
<tr>
<td></td>
<td>all+global(asr)</td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td>65.2</td>
<td>58.0</td>
</tr>
<tr>
<td>Recall</td>
<td>65.5</td>
<td>61.6</td>
</tr>
</tbody>
</table>

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 and linguistic features extracted automatically (all(asr)) improves performances. In comparison to the use of linguistic features coming from manual transcriptions, by merging acoustic and linguistic features extracted from the ASR outputs, we obtain better results whatever the class of spontaneity or the metric used, except in terms of recall for prepared speech.
By examining the results of the global+all(asr) condition, we observe that the probabilistic contextual tag model applied on the all(asr) condition allows to significantly improve the performance of the classification whatever the class of spontaneity or the metric used.

In fact, in this article we are particularly interested on the detection of high spontaneous speech segments. By accepting all the propositions of the classification, our method allows to achieve a 69.3% precision for high spontaneous speech detection with a 74.6% recall measure, as presented in Table II. More precisely, 83.5% of high spontaneous detection errors are due to confusion between low and high spontaneous speech.

Using the scores $c(s_i)$ given by the classifier combined with the probabilities $P(s_i|s_{i-1}, s_{i+1})$ provided by the contextual tag model, it is possible to filter the proposition by applying a threshold to the value of $c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$.

Figure 1 presents the detection performance obtained by changing the threshold on classification score for high spontaneous segments: we can see that our system could be more accurate (precision increase) when we take less decisions (recall decrease). This possibility of thresholding can adapt the use of the classification method by finding the best compromise between recall and precision for the targeted application.

The new global approach all+global(asr) using a probabilistic model really outperforms the previous local approach all(asr): whatever the threshold value, this new approach allows to reach a better precision with a better recall.

VI. 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 Broadcast News) has been manually labelled according to a level of spontaneity.

The acoustic and linguistic features are evaluated in order to characterize and detect spontaneous speech segments: the combination of acoustic and linguistic features extracted from ASR outputs obtains a better precision and a better recall than the linguistic features extracted from the reference transcriptions alone. Moreover, using a probabilistic contextual tag-sequence model to globalize the classification process allows a better 74.6% precision in the detection of high spontaneous speech with a 69.3% recall measure, and 83.5% of high spontaneous speech detection errors are due to confusion between low and high spontaneous speech.

By applying a threshold on the scores obtained during the classification process, the high spontaneous speech detection precision can reach 85%, but with a recall equals to 25%. Although the classification task of labeling speech segments according to the spontaneity level is hard — even for human annotators — much progress has been made in the automatic detection since our previous work [1], mainly due to an improvement of the ASR system, but especially due to the addition of the probabilistic contextual tag model allowing a more global classification process.

This spontaneous speech detection provides very useful piece of information which can be used by various applications: speech recognition, for example, by developing specific methods for minimizing the word error rate on this kind of speech; but also, for example, by providing additional information to automatic structuring or classification of collections of audio documents in large audio database.

REFERENCES