A Language-identification inspired method for spontaneous speech detection

Mickael Rouvier, Richard Dufour, Georges Linarès and Yannick Estève

1CERI-LIA, University of Avignon
2LIUM, University of Le Mans

{mickael.rouvier, georges.linares}@univ-avignon.fr
{richard.dufour, yannick.esteve}@lium.univ-lemans.fr

Abstract

Most of spontaneous speech detection systems relies on disfluency analysis or on combination of acoustic and linguistic features. This paper presents a method that considers spontaneous speech as a specific language, which could be identified by using language-recognition methods, such as shifted delta cepstrum parameters, dimensionality reduction by linear discriminant analysis and factor-analysis based filtering process. Experiments are conducted on the French EPAC corpus. On a 3 spontaneity-level task, this approach obtains a relative gain of about 22% of identification rates, in comparison to the classical MFCC/GMM technique.

Then, we combine these techniques to others previously proposed for spontaneous speech detection. Finally, the proposed system obtains a recognition rate of 65% on high spontaneous speech segments.

Index Terms: speech processing, spontaneous speech detection

1. Introduction

Spontaneous speech detection issue was pointed out by numerous authors in the field of speech recognition and understanding. One of the main difficulty lies in the fact that linguistic and acoustic structures of spontaneous speech completely differ from read or prepared speech: speakers frequently hesitate, interrupt themselves, change their speaking rate, etc. Previous works investigate detection and correction of disfluencies [1]. In [2], authors proposed to use specific linguistic and prosodic descriptors extracted from an automatic speech recognition (ASR) system to determine the level of spontaneity of each speech segment. More, they propose to take into consideration information about surrounding segments by using the nature of the contiguous neighboring speech segments.

Here, the basic idea is that spontaneous and prepared speech could be distinguished in the same way as different dialects or languages. Following this way, we propose to combine language-identification (LID) inspired methods and previous method [2], based on linguistic and structural descriptors, yet successfully used for spontaneous speech detection.

Next section describes the experimental setup. Then, we present the global architecture overview of the proposed system. Section 4 describes the acoustic features for spontaneity level estimate. Combination of acoustic and linguistic features is described in Section 6. Last section concludes and presents some perspectives.

This research was supported by the ANR agency (Agence Nationale de la Recherche), RPM2 project (ANR-07-AM-008)

2. Corpus and Task

Experiments are conducted on the EPAC French corpus, composed of spontaneous parts from Broadcast News. Each speech segment is annotated with a set of ten labels, each corresponding to a spontaneity level: grade 1 stands for prepared speech, almost similar to read speech, and grade 10 stands for very disfluent speech, almost non understandable. In our experiments, 3 classes are considered: prepared speech (E1) corresponding to grade 1; low spontaneity (E2) corresponding to grades 2 to 4; and high spontaneity (E3) corresponding to grades 5 and over.

The total duration of speech materials is of 11h37 for 3322 speech segments. 1142 of these segments are labeled as prepared speech, 1175 as low spontaneity and 1005 as high spontaneity. 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 are evaluated.

3. System overview

The proposed system is composed of 2 levels, performing a local and a global evaluation of the spontaneity level.

The first level consists in features extraction. We identify 2 independent methods of features extraction, respectively corresponding to acoustic (LID) and prosodic/linguistic descriptors. The first method uses acoustic features firstly estimated on short term cepstral coefficients. GMM classifier estimates the probability of classes (prepared, low and high spontaneous speech) knowing each acoustic observation. The resulting frame-level scores are then cumulated to evaluate the classification hypothesis on the whole observation sequence. The second method relies on linguistic and prosodic features extracted from an automatic speech recognition (ASR) system [6]. These features are related to linguistic and prosodic analysis of spoken contents. They are combined by using a classical boosting algorithm.

The level 2 performs a general estimation by integrating these locally-estimated probabilities into a global model [2]. This model rescores probabilities of segment-level spontaneity according to the context in which the segment occurs.

4. Acoustic features

Recent LID systems combine SDC-MFCC features and variability reduction by Factor Analysis (FA). In the next section, we compare MFCC and SDC-MFCC features for spontaneous speech detection and we evaluate HLDA-based dimensionality reduction, that was largely applied to speech processing methods. Finally, we apply FA method in a similar way that it was successfully used on speaker and language identification tasks.
4.1. Shifted Delta Cepstrum (SDC)

The SDC feature was originally proposed in [3] for language identification. The purpose is to capture cepstral dynamics on a larger temporal window.

The computation of the SDC features is illustrated in figure 1. The SDC features are determined by 4 parameters: \( N, d, P \) and \( k \), where \( N \) is the number of cepstral coefficients computed at each frame, \( d \) represents the time delay for the delta computation, \( k \) is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and \( P \) is the time shift between consecutive blocks. Accordingly, \( kN \) parameters are used for each SDC feature vector, as compared with \( 2N \) for conventional cepstra and delta-cepstra feature vectors. For example, the vector at frame time \( t \) is given by the concatenation of all the \( c(t + iP) \), where:

\[
\Delta c(t) = c(t + iP + d) - c(t + iP - d) \quad (1)
\]

Figure 1: Computation scheme of Shift Delta Cepstrum.

We compare MFCC and SDC-MFCC on a GMM classifier. For both, we used 256 mixture models trained by likelihood maximization with expectation-maximization algorithm. Results are reported in table 1.

Table 1: F-measure, precision and recall depending on class of spontaneity, by GMM-classification in cepstral domain (with MFCC and SDC-MFCC techniques).

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.46</td>
<td>0.28</td>
<td>0.42</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>(0.52/0.42)</td>
<td>(0.25/0.33)</td>
<td>(0.42/0.42)</td>
<td></td>
</tr>
<tr>
<td>SDC-MFCC</td>
<td>0.47</td>
<td>0.28</td>
<td>0.48</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>(0.50/0.44)</td>
<td>(0.24/0.33)</td>
<td>(0.50/0.44)</td>
<td></td>
</tr>
</tbody>
</table>

Results show that SDC-MFCC slightly outperforms MFCC, classification rates increasing from 39% to 41%. We note that SDC-MFCC seems especially efficient on highly spontaneous speech (E3): we obtained 42% and 48% of F-Measure, respectively for MFCC and SDC-MFCC.

4.2. Factor Analysis for spontaneous speech characterization

Gaussian Mixture Models (GMM) and GMM-UBM (Universal Background Model) constitute a classical approach to speaker verification. Here, the world model (UBM-GMM) represents a generic speech model trained on the whole training corpus. The 3 spontaneity-level-specific GMMs comes from a world model adaptation, achieved by the classical MAP (Maximum A Posteriori) adaptation technique. Only mean vectors are adapted, weights and variances remaining constant with respect to those of the UBM. FA is used to decompose the spontaneous-speech-specific model into three different components: a generic-speech component, a spontaneous-speech-dependent component and a session-dependent component (each recording corresponding to one session). A GMM mean supervector is defined as the concatenation of the means of the gaussian components of the GMM. Let \( D \) be the dimension of the feature space, the dimension of a supervector mean is \( MD \), where \( M \) is the number of components in the GMM. For a spontaneous speech (SS) belonging in session \( h \), the FA model can be formulated as:

\[
m_{(h,SS)} = m + Dy + Ux_{(h,SS)}, \quad (2)
\]

where \( m_{(h,SS)} \) is the session-dependent supervector mean, \( D \) is \( MD \times MD \) the diagonal matrix, \( y \) the observation vector (a MD vector), \( U \) the session variability matrix of low rank \( R \) (a \( MD \times R \) matrix), and \( x_{(h,SS)} \) the channel factors (a \( R \) vector). All parameters of the FA model are estimated by using Maximum Likelihood criterion and algorithm EM. Several sessions corresponding to each spontaneity grade have to be used for accurate estimate of FA parameters. This compensated model is simply obtained from equation:

\[
m_{SS} = m + Dy, \quad (3)
\]

As reported in table 2, performance is strongly improved by FA. On SDC-MFCC based system, the relative error rate reduction is about 21%, absolute identification rates increasing from 41% to 53%.

Table 2: F-measures, precision and recall depending on class of spontaneity classification using factor analysis.

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.47</td>
<td>0.28</td>
<td>0.42</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>(0.52/0.44)</td>
<td>(0.25/0.33)</td>
<td>(0.42/0.42)</td>
<td></td>
</tr>
<tr>
<td>SDC-MFCC</td>
<td>0.47</td>
<td>0.28</td>
<td>0.48</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>(0.50/0.44)</td>
<td>(0.24/0.33)</td>
<td>(0.50/0.44)</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Heteroscedastic Linear Discriminant Analysis (HLDA)

HLDA has been used on LID (and on other speech processing systems [4]) to improve class separability and reduce acoustic space dimensionality [5]. Here, our classifier operates on SDC features which comes from a grouping of close frames. One of the potential drawback of this approach is due to redundancies which increasing the representation space and could negatively impact the system accuracy.

Figure 2: Classification rate according to the HLDA order.
We apply HLDA to the previously presented SDC-MFCC features, and we evaluate the impact of dimensionality reduction on system accuracy. Results reported in figure 2 confirms the interest of HLDA. With the best configuration, we obtain an absolute gain of about 2% on classification rate with a slight reduction of acoustic space dimensionality (from 56 to 50).

5. Prosodic and linguistic based approach

As described in [6, 2], another approach is to extract some prosodic and linguistic features from ASR system in order to distinguish prepared and spontaneous speech. Experiments are made on the corpus presented in section 2.

5.1. Prosodic features

Particularities of vowel duration and phonetic rates are relevant indexes of spontaneous speech. Indeed, some specific studies [7, 8] shown that vowel duration, the lengthening of a syllable at the end of a word, or the concept of melism are useful features to define this particular kind of speech. Regarding the phonetic rate, [8] have shown the correlation between the variations of speech rate and the emotional state of a speaker. So an estimate of the speech rate, by speech segment, is another interesting prosodic feature that will be extracted. Finally, the extraction of pitch is the last prosodic feature used.

5.2. Linguistic features

Speech disfluences, categorized as filled pause, repetition, repair and false start, have been described in many work [7, 9] to characterize spontaneous speech. The idea here is to extract from transcriptions of an ASR system these disfluences, to use them as linguistic features. Filled pauses (with its number of occurrences) and repetitions and false starts (with its counts of 1-gram and 2-gram) will be estimated from ASR output.

Moreover, [10] shown that agrammaticity and different language register can define spontaneous speech. To get these phenomena, a POS tagging and a syntactic chunking process is realized on ASR outputs to extract some features like bags of n-gram, POS tags and syntactic chunk categories, including their count and average length.

Later, the number of occurrences of proper nouns in a speech segment is used, as analyzed by [11]. Average and variance of confidence measures provided by an ASR system constitute our last feature.

5.3. Classification

The classification process takes into consideration the acoustic and linguistic descriptors presented in section 5.1 and 5.2, plus some other descriptors as the duration of speech segments and the number of recognized words. It proposes a categorization of the speech segments according the three class of spontaneity. Each segment is processed individually.

The classification tool used is *icistoost*, 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*.

5.4. Probabilistic model for global decision

The approach, described in [2], shows 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. A classical statistical approach is proposed, using a maximum likelihood method.

Let be $s_i$ a tag of the segment $i$, with $s_i \in \text{high spontaneity}$, low spontaneity, prepared. We define $P(s_i \mid s_{i-1} = \text{high spontaneity}, s_{i+1} = \text{low spontaneity})$ as the probability of observing a segment $i$ associated to the $s_i$ 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 icistoost classifier on choosing the tag $s_i$ for the speech segment $i$ according to the values of the descriptors extracted from this segment. $\hat{S}$ 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 $S$ which maximizes the global score obtained by combining $c(s_i)$ and $P(s_i \mid s_{i-1}, s_{i+1})$ for each speech segment $i$ detected on the audio file. The sequence $\hat{S}$ is computed by using the following formula:

$$\hat{S} = \arg\max_{\hat{S}} c(s_1) \ast c(s_n) \ast \prod_{i=2}^{n-1} c(s_i) \ast P(s_i \mid s_{i-1}, s_{i+1})$$

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

Table 3 presents results obtained with boosting classification method using ASR extracted features at segment level *L+A (local)* and at global level *L+A (global)*. Results shows that using a global method improves performance obtained at segment level, with an improvement of detection performance on every class of spontaneity, and with a total gain of 8 absolute points.

### Table 3: F-measures, recall and precision according to the three classes of spontaneity, using ASR features extracted at segment level (local) and at global level.

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>L+A (local)</td>
<td>0.63</td>
<td>0.55</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>(0.660/0.61)</td>
<td>(0.430/0.47)</td>
<td>(0.660/0.55)</td>
<td></td>
</tr>
<tr>
<td>L+A (global)</td>
<td>0.68</td>
<td>0.51</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>(0.660/0.67)</td>
<td>(0.300/0.32)</td>
<td>(0.710/0.71)</td>
<td></td>
</tr>
</tbody>
</table>

6. System combination

Methods previously presented can already reach high levels of performance on kind of speech detection. Thereby, the method using LID approaches and the method using prosodic and linguistic features extracted from an ASR system (without using the global method) obtain separately almost identical performance on the global classification rate (respectively 0.55 and 0.56 in term of F-measure).

For now, the approach chosen for both statistical methods is to get the class of spontaneity associated with best confidence score computed. So combining methods could be a solution to still improve detection performance. As many combinations are possible, we focused on using confidence scores provided by the best LID approach, which will be integrated in various components of method proposed in [2] (local and global methods). We study three combinations using these confidence measures:

- using them as input of probabilistic model for global decision *LID (global)*.
- including them as new features of classification method using the AdaBoost algorithm *L+A incl. LID*.
- using confidence scores of *L+A incl. LID* combination as input of global decision *L+A incl. LID (global)*.
Table 4 presents results obtained with the three combinations, depending on each class of spontaneity. The global classification rate (Total) computed on all classes of spontaneity is also presented. When focusing on LID (global), gains are particularly visible on high spontaneous class (E3), where recall is higher than any other method. This first combination improves also the global classification rate (Total), with a gain of 4 absolute points comparing to the best result with standalone LID method. We can see that the second combination L+A incl. LID gives interesting results, as the global classification rate is higher than L+A (local). Finally, the last combination L+A incl. LID (global) seems to give the best general results: the global classification rate, F-measures on E1 and E2 classes, and the precision on E3 are better than all other proposed methods. The only decrease can be seen on recall of E3, which is very high with the LID (global) combination method, but with a low precision.

As confidence measures give an estimation for each class of spontaneity and segment, a threshold could be done to increase detection precision. As many tasks have to focus their efforts on spontaneous speech, figure 3 presents detection performance of best standalone and combination detection methods presented bellow, by varying the threshold on confident scores on this specific kind of speech.

![Detection performance on classification techniques of E3 segments according to a varying threshold on the confidence scores.](image)

As expected, the combination method L+A incl. LID (global) outperforms other methods in term of precision and recall when varying threshold (except for recall on LID (global) combination).

7. Conclusions

We integrated new acoustic-based techniques in a previously presented framework for spontaneous speech detection. These techniques come from the field of language identification, where specific approaches have been developed for extra-linguistic speech characterization, such as medium-term cepstral dynamics and intra-class variability reduction. Our results demonstrate the effectiveness of this approach: we observed a gain of about 16% in terms of recognition rates, on the task of classification of speech segments according 3 levels of spontaneity. Moreover, the low-level acoustic features evaluated do not require any ASR module, that could be dramatically impacted on adverse or unexpected conditions.

These results show that low-level characteristics of spontaneous speech present some similarities with foreign languages. Nevertheless, the analogy between spontaneous speech and foreign languages seems not so relevant from an higher-level view: phonetic and linguistic systems, that may be considered as discriminative features for languages, could not be effective to distinguish prepared and spontaneous speech. This question should be investigated in the future. Moreover, we plan to evaluate other low-level acoustic features that are used for emotion recognition.

8. References


