Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 1–18

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Characterizing and detecting spontaneous speech: Application to speaker role recognition

Richard Dufour *, Yannick Estève, Paul Deléglise

LIUM, University of Le Mans, France

Received 11 December 2012; received in revised form 12 July 2013; accepted 27 July 2013
Available online 7 August 2013

Abstract

Processing spontaneous speech is one of the many challenges that automatic speech recognition systems have to deal with. The main characteristics of this kind of speech are disfluencies (filled pause, repetition, false start, etc.) and many studies have focused on their detection and correction. *Spontaneous speech* is defined in opposition to *prepared speech*, where utterances contain well-formed sentences close to those found in written documents.

Acoustic and linguistic features made available by the use of an automatic speech recognition system are proposed to characterize and detect spontaneous speech segments from large audio databases. To better define this notion of spontaneous speech, segments of an 11-hour corpus (French Broadcast News) had been manually labeled according to three classes of spontaneity.

Firstly, we present a study of these features. We then propose a two-level strategy to automatically assign a class of spontaneity to each speech segment. The proposed system reaches a 73.0% precision and a 73.5% recall on high spontaneous speech segments, and a 66.8% precision and a 69.6% recall on prepared speech segments.

A quantitative study shows that the classes of spontaneity are useful information to characterize the speaker roles. This is confirmed by extending the speech spontaneity characterization approach to build an efficient automatic speaker role recognition system.

Keywords: Spontaneous speech; Speaker role; Feature extraction; Speech classification; Automatic speech recognition; Role recognition

1. Introduction

Extracting information from large audio databases becomes a very challenging task since the amount of available audio data continues to grow (for example podcast, online video-sharing, etc.). The extraction of the audio document structure as well as the linguistic content is needed to retrieve high-level information. For example, a part of this information retrieval process is adding sentence punctuation and boundaries in automatic transcriptions. This segmentation process is very important for many tasks such as speech summarization, speech-to-speech translation or the distillation task as defined in the GALE program in Hakkani-Tür and Tür (2007). Nevertheless, the difficulty of providing this structure depends on other phenomena, such as the type of speech. Indeed, this process is much more difficult in the presence of spontaneous speech, as this kind of speech is characterized by ungrammaticality and disfluencies. Moreover, in order to cluster some documents according to their content or structure, the presence of spontaneous speech segments should be an interesting descriptor. It is therefore useful to characterize the spontaneity level of speech segments at an early stage in order to adapt automatic speech recognition (ASR) systems, as presented in Dufour et al. (2010a).

The type of speech inside broadcast audio data can switch between parts of prepared speech (news presentations, reports, etc.) and more spontaneous ones (interviews,

* Corresponding author. Permanent address: LIA, University of Avignon, France. Tel.: +33 490843502.

E-mail addresses: richard.dufour@univ-avignon.fr (R. Dufour), yannick.esteve@lium.univ-lemans.fr (Y. Estève), paul.deleglise@lium.univ-lemans.fr (P. Deléglise).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.07.007

debates, dialogues, etc.). The main characteristics of spontaneous speech are disfluencies (filled pause, repetition, repair and false start), and many studies have focused on their detection and their correction (Goto et al., 1999; Liu et al., 2005; Lease et al., 2006) as pointed out by the NIST Rich Transcription Fall 2004 evaluation. All these studies show an important drop in performance between results obtained on reference transcriptions and those obtained on automatic transcriptions. This could be explained by the noise generated by the ASR systems on spontaneous speech *segments*, which produce higher Word Error Rates (WER) than those obtained on *prepared* speech. A *segment* refers to a portion of audio signal in an audio file. Segments can contain speech, music, etc. In speech recognition, automatic segmentation splits audio documents into segments which last between 5 and 20 s. These segments may contain long pauses which generally define their boundaries.

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, as shown in Boula de Mareüil et al. (2005). 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. Prepared speech is produced by speakers who have enough time to prepare their intervention.

We propose to consider a set of acoustic and linguistic features for characterizing *spontaneous speech* limited to features that can be extracted from an automatic speech recognition processing only. This choice was motivated by the availability of these features when audio documents are indexed from their lexical content thanks to automatic transcriptions. The relevance of these features is estimated on an 11-hour corpus (French Broadcast News) manually labeled according to 3 classes of spontaneity. We then propose a speech spontaneity characterization system which will automatically assign a class of spontaneity to each speech segment. Indeed, this method involves two major approaches:

- **Local process**: individual classification of each speech segment according to its class of spontaneity using the set of acoustic and linguistic features studied.
- **Global decision**: the nature of the contiguous neighboring speech segments is taken into account. Thereby, the categorization of each speech segment has an impact on the categorization of the other ones.

We also propose to apply our automatic speech spontaneity characterization system on a manually labeled speaker role corpus to assess our speech spontaneity detection method on a new task to see if a correlation exists between a speaker role and a class of spontaneity. Then, we propose to directly use this speech spontaneity characterization method to recognize speaker roles.

This article is an extension of a previous work presented in Dufour et al. (2009a, 2011), and provides more details about our speech spontaneity detection method based on acoustic and linguistic features made available during a speech recognition process. More, in this article, a fully automatic speaker role recognition system is presented with experimental results. Section 2 presents a study of the extracted features to characterize the level of spontaneity of each speech segment. We also describe the correlation between the Word-Error-Rate obtained by a state-of-the-art ASR decoder on this broadcast news corpus and the level of spontaneity. We then propose, in Section 3, a two-step automatic speech spontaneity characterization system: the first step individually classifies each speech segment with a class of spontaneity, while the second one takes advantage of a global decision process to improve the spontaneity speech characterization. A study of the speech spontaneity and speaker role relationship is presented in Section 4. The speaker role recognition system based on our speech spontaneity detection method is finally presented in Section 4.4.

2. Spontaneous speech characterization

2.1. Levels of spontaneity

By defining spontaneous speech as *unprepared speech*, it is possible to follow a definition proposed by Luzzati (2004) that defined a spontaneous utterance as: “a statement conceived and perceived during its utterance”. This definition illustrates the classification subjectivity between prepared and spontaneous speech. Ideally, to annotate a speech corpus with labels representing the fluency of each speech segment, each speaker would have to annotate his own utterances. As this seems not feasible, an annotation protocol had been established. The protocol is based on a human judge perception using a *level of spontaneity* for a given speech segment. Our approach was to manually tag a corpus of speech segments with a set of 3 labels, each one corresponding to a spontaneity level: from grade 1, which stands for prepared speech, almost similar to read speech, to grade 3, which applies to very disfluent speech, almost not understandable.

Two human judges annotated a speech corpus by listening to the audio recordings. The corpus had been cut into segments using a state-of-art automatic segmentation and diarization process proposed in Meignier and Merlin (2010). The segmentation system has a diarization error rate (DER) between 5.71% on the EPAC corpus, as presented in Estève et al. (2010), and 10.01% on the ESTER 2 corpus, as shown in Meignier and Merlin (2010). No transcriptions were provided to the annotators. In order to evaluate inter-annotator agreement for this specific annotation task on the 3 grades presented above, we computed the Kappa coefficient of agreement (Cohen, 1960) on one hour of broadcast news separately annotated by the two human judges. The coefficient obtained was very high:

0.852 — a value greater than 0.8 is usually considered as excellent (Eugenio and Glass, 2004).

Then, the manual annotation process continued with the labeling of the ten remaining hours. Since each judge labels all the data, some annotation differences could appear; in this case, human annotators together find a consensus.

Moreover, one of the problems encountered was that spontaneous speech segments can occur anywhere, not only in conversational speech, but also in the middle of very *clean* utterances. Similarly, even conversational speech can contain segments which could be considered as prepared speech. To take this specificity into account, we decided to independently evaluate each segment: a spontaneous segment can be surrounded by many prepared ones.

The corpus obtained after this labeling process is made of 11 files containing French Broadcast News data from 5 different French radio stations (France Culture, France Inter, France Info, Radio Classique and RFI). The total duration is 11 h with a total of 3,814 segments (after removal of the non speech segments: music, jingles, etc.). Among these segments, 1,228 were annotated with the *prepared speech* label, 1,339 with the *low spontaneous* label and 1,247 with the *high spontaneous* label.

This approach allows us to subjectively choose the limit between spontaneous and prepared speech. In these experiments, we considered 3 classes:

- *Prepared* speech corresponds to grade 1.
- *Low spontaneous* speech corresponds to grade 2.
- *High spontaneous* corresponds to grade 3.

This corpus and its annotation process has already been presented in Dufour et al. (2009a,b, 2010b). It is publicly available with the EPAC data through the European Language Resources Association (ELRA) catalog for research or commercial use.[1]

2.2. Acoustic and linguistic features

In addition to the subjective annotated corpus presented in the previous section, we now introduce a study presenting some particularities of spontaneous speech. We chose to describe features extracted from speech segments that are relevant to characterize speech spontaneity. This problem has been studied as a specific task of the Rich Transcription Fall 2004 blind evaluation, which focused on the detection of speech disfluencies. It is interesting to note that proposed methods to detect them use various sources of information: linguistic features (Lease et al., 2006), both linguistic and prosodic features (Liu et al., 2006), or linguistic and more generic acoustic features (Yeh and Wu, 2006).

These works led us to study three feature sets: acoustic features related to prosody, linguistic features related to the lexical and syntactic content of the speech segments, and finally ASR confidence measures. The objective of this study is to highlight specific features that could be embedded into an automatic classification process to assign a class of spontaneity to each 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. We chose to focus our work on features already studied as specificities of the spontaneous speech and that could be extracted by an ASR system. Indeed, we did not studied all extractable features contained in the speech signal, such as those proposed in the Interspeech Challenges (Schuller et al., 2012), which could constitute an interesting perspective of this work. The extracted features and their quantitative study are presented in the next section.

2.2.1. Prosodic features

The prosodic features used are related to vowel duration and phonetic rate:

- **Duration:** following previous work describing the link between prosody and spontaneous speech presented in Shriberg (1999), we chose to study two features: vowel duration and syllable lengthening at the end of a word. The latter had been proposed in Caelen-Haumont (2002) 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 (Caelen-Haumont, 2002) have also shown the correlation between the speech rate variations and the emotional state of a speaker. Following this idea, we use, as a feature, an estimate of the speech rate by speech segment, in order to observe its impact on the speech spontaneity. We decided to estimate the phonetic rate in two ways: the average and the standard deviation of the phone duration on the whole segment, firstly by considering pauses and fillers as phones, and secondly by removing them. As presented in Dufour et al. (2009b), we also study the fundamental frequency f0 (*Pitch*) standard deviation as a potential feature. This feature is rarely used for ASR, but it was interesting to study another kind of feature in association with the chosen ASR-based features.

To extract phonetic rate and duration features, a forced alignment had been performed to align the phonemes contained in transcription data to the speech signal. As a result, we got the duration of each phoneme contained in each word. The alignment had been performed using the ASR system presented in 3.2. Finally, pitch values had been extracted with the *Snack Sound Toolkit*.[2]

[1] Reference ELRA-S0305: http://catalog.elra.info/product_info.php?products_id=1119

[2] http://www.speech.kth.se/snack/

Table 1
Comparison of average duration (in ms) of vowels, phone duration with (w/) and without (w/o) fillers, and melisms according to the three classes of spontaneity.

Features	Average			Standard deviation		
	Prepa.	Low	High	Prepa.	Low	High
Vowel duration	75	81	91	41	53	75
Phone duration w/ fillers	81	85	92	46	57	73
Phone duration w/o fillers	78	81	87	40	47	62
Melisms	82	94	110	45	63	92
Pitch (Hz)	–	–	–	32.43	32.65	32.93

Table 1 presents the average on duration and the average on standard deviation of vowel duration, phone duration (including vowels) with (w/) and without (w/o) fillers, and melisms depending on the 3 levels of spontaneity. Average of standard deviation of pitch are also presented. All these values were computed on the experimental data described in Section 2.1.

Results show that a correlation exists between these features and the level of spontaneity. Indeed, average durations and standard deviations of the acoustic features are higher on *high spontaneous* speech segments.

2.2.2. Linguistic features

In parallel, we studied some linguistic features by focusing on the syntactic and lexical segment content. The concept of *speech disfluencies* is the main characteristic of spontaneous speech. Disfluencies can be categorized as filler word (Duez, 1982), repetition, repair (Heeman et al., 1996) and false start (O'Shaughnessy, 1993). Various studies focused to describe them at the acoustic (Shriberg, 1999) or lexical level (Siu and Ostendorf, 1996).

Thereby, we choose to analyze two linguistic features, which was supposed to represent the speech segment spontaneity level:

- **Filler words**: the ASR lexicon contains several filler words in French. For example, we could have *euh*, *ben* or *hum* inside the final ASR transcription. The total number of filler word occurrences encountered in a segment divided by the total number of words constitutes the first linguistic feature.
- **Repetition**: the second feature investigates the problem of repeated words by dividing the number of unigram and bigram repetitions in a segment by the total number of words.

As shown by Boula de Mareüil et al. (2005) on broadcast news data, spontaneous speech can also be characterized at the linguistic level by other phenomena. Indeed, ungrammaticality and language register are also very typical of unprepared speech. In order to catch the link between spontaneity on one side and lexicon and syntax on the other side, we apply to the transcripts a shallow parsing process including a Part-Of-Speech (POS) tagging and a syntactic chunking process. This linguistic information had been extracted with the *Lia_tagg* tool.[3] We chose to use the following features, computed for each speech segment:

- **Bags of n-grams** (from 1 to 3-g) on words, POS tags, and syntactic chunk categories (noun phrase, prepositional group).
- **Average count** of syntactic chunks, words, and POS tags count.

Moreover, following the experiments carried out by Bazillon et al. (2008), we enriched the linguistic features with the number of proper noun occurrences in a speech segment divided by the total number of words. Indeed, authors showed that a higher number of proper nouns appears in context of *prepared speech*, in comparison with *spontaneous speech*.

In order to better understand the impact of linguistic features, we compare results obtained for each of them depending on the speech spontaneity level. Thus, Table 2 presents the average scores obtained by each linguistic feature for each class of spontaneity using the reference transcriptions and the ones computed from the automatic transcriptions provided by the ASR system.

2.2.3. ASR confidence measures

Confidence measures are computed scores which express reliability of recognition decisions made by an automatic speech recognition system. These scores could be used to characterize speech segment spontaneity. As already discussed in Section 1, ASR systems have more difficulties to correctly transcribe spontaneous speech than prepared speech. Table 3 presents the mean and the standard deviation of the ASR confidence measure for each class of spontaneity obtained with the labeled corpus. These confidence measures were provided by our ASR system, described in Section 3.2. They are *a posteriori* probabilities computed from confusion networks. These confusion networks are built from word-graphs containing acoustic and linguistic scores for each word. For each segment, we compute the average value of ASR confidence measures associated to each word.

[3] http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html

Table 2
Comparison of average values for each linguistic feature depending of the class of spontaneity on reference transcriptions. Values are in proportions per segment for filler words, repetition and proper nouns, and number of words for chunk size.

Features	Reference trans.			Automatic trans.		
	Prepa.	Low	High	Prepa.	Low	High
Filler word (%)	0.30	2.13	4.51	0.21	1.63	3.81
Repetition (%)	0.14	0.65	3.15	0.14	0.51	2.24
Proper noun (%)	8.33	5.61	2.99	7.39	5.33	3.45
Chunk size (# words)	1.73	1.67	1.57	1.71	1.66	1.56

Table 3
Comparison of the confidence measure average and standard deviation for each speech category.

	Prepared	Low spontaneous	High spontaneous
Average	0.91	0.88	0.82
Standard deviation	0.133	0.152	0.183

A gradation of results can be observed, depending on the level of spontaneity: the average values decrease when speech becomes more spontaneous, while standard deviation values tend to increase. ASR confidence measures seem to be a good indicator of speech spontaneity. Thus, confidence measures will be associated with acoustic and linguistic features to improve the classification of speech segments with the three classes of spontaneity. We proposed to use this feature in Dufour et al. (2010b).

3. Automatic detection of spontaneous speech segments

3.1. General approach

We propose to treat type of the speech detection as a multiclass classification problem. The main idea is to combine all the extracted features (acoustic, linguistic and ASR confidence measures) in order to label each speech segment with a class of spontaneity, as presented in Dufour et al. (2009b). This combination will be made through a classification process, which will provide the most likely class of spontaneity based on extracted features. Fig. 1 summarizes the followed approach to assign a class of spontaneity to each speech segment.

Since this approach only uses information at the segment level, we suspected that the decision should be made at a higher level. Indeed, it should be rare to observe a speech segment labeled as *high spontaneous* speech surrounded by segments labeled as *prepared* speech. In order to improve our approach, we propose to take into account the nature of the contiguous 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. This approach has been presented in Dufour et al. (2009a). We chose to use a statistical classical approach by using a maximum likelihood method. Fig. 2 presents the global decision process, using results obtained at the segment level.

3.2. The LIUM speech transcription system

To automatically extract the acoustic, linguistic and confidence measure descriptors for categorizing speech segments according to the class of spontaneity, we used the LIUM ASR system described in Deléglise et al. (2009). The LIUM ASR system had been developed in order to participate in the French ESTER 2 evaluation campaign on French Broadcast News automatic transcription systems. This system also participated to more recent campaigns: it reached the first rank in 2012 in the French ETAPE benchmark on speech-based TV content (Gravier et al., 2012) and was used to process English language in the LIUM spoken language translation system which won the TED talks task during the IWSLT'11 international campaign (Rousseau et al., 2011).

This ASR system includes a state-of-the-art speaker diarization system also developed at LIUM (Meignier and Merlin, 2010). This speaker diarization tool obtained the best diarization error rate during the ESTER 2 evaluation campaign on december 2008 and more recent campaigns as ETAPE, and is now used under an open-source license in many laboratories all over the world. The LIUM automatic speech recognition system is based on the CMU Sphinx system. The tools distributed in the CMU Sphinx open-source package,[4] although already at a high level of quality, can be supplemented or improved to integrate some state-of-art technologies. This is the solution adopted by the LIUM to develop its ASR system: gradually extending the Sphinx ASR tools and bringing it to new performance levels (Deléglise et al., 2009). For this work, the acoustic and language models of the ASR system were estimated on the ESTER 2 data. Details of this ASR system can be found in Deléglise et al. (2009)

3.3. Classification at the segment level

The features presented in Section 2.2 are evaluated on our labeled corpus with a multiclass classification task. The classification tool used is *icsiboost*,[5] an open-source tool based on the *AdaBoost* algorithm such as the *Boostexter* software (Schapire and Singer, 2000). 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 (useful for linguistic features) or a numerical value (discrete or continuous: useful for acoustic, linguistic and confidence measure features). At the end of the training process, a list of weighted rules is obtained. With this set of rules, a score for each class is computed on each data to classify.

[4] http://cmusphinx.sourceforge.net/
[5] http://code.google.com/p/icsiboost/

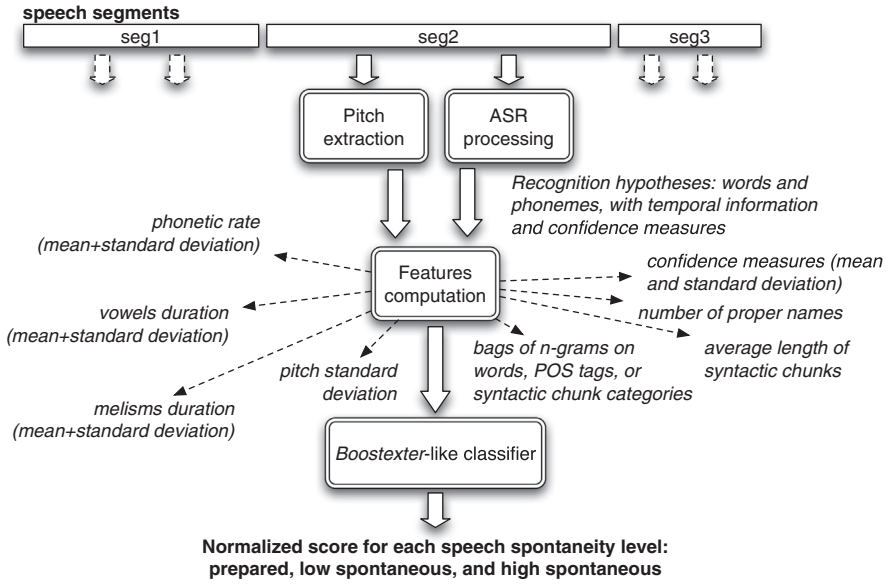

Fig. 1. General approach for the classification process involving a class of spontaneity at each speech segment.

This classification process proposes a categorization of the speech segments according to the three classes of spontaneity (*prepared*, *low spontaneous* or *high spontaneous* labels). It takes into consideration the acoustic and linguistic descriptors, the ASR confidence scores, and additional decoding information, such as the duration of each segment and the total number of recognized words. Each segment is processed individually.

3.4. Probabilistic contextual model for global decision

3.4.1. General approach

Let s_i be a tag of the segment i, with $s_i \in \{$"high spontaneous", "low spontaneity", "prepared"$\}$. We define $P(s_i|s_{i-1}, s_{i+1})$ as the probability of observing a segment i tagged as s_i when the previous segment is tagged as s_{i-1} and the next segment is tagged as s_{i+1}. Let $c(s_i)$ be the normalized confidence measure given by the *AdaBoost* classifier when choosing the tag s_i for the speech segment i according to the values of the descriptors extracted from this segment. We have:

$$\sum_{s_i} c(s_i) = 1$$

S is a sequence of tags s_i assigned to the sequence of all the speech segments i (only one tag per segment). The global decision process consists in choosing the tag-sequence hypothesis $\bar{S}$ 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 $\bar{S}$ is computed by using the following equation:

$$\bar{S} = \underset{S}{\operatorname{argmax}}\, c(s_1) \times c(s_n) \times \prod_{i=2}^{n-1} \times c(s_i) \times P(s_i|s_{i-1}, s_{i+1}) \quad (1)$$

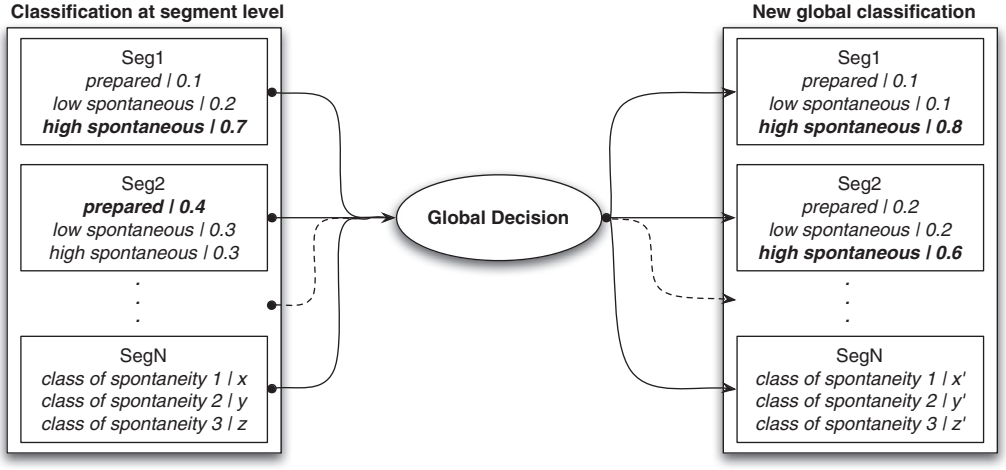

Fig. 2. General approach of the global decision process for labeling speech segments with a class of spontaneity.

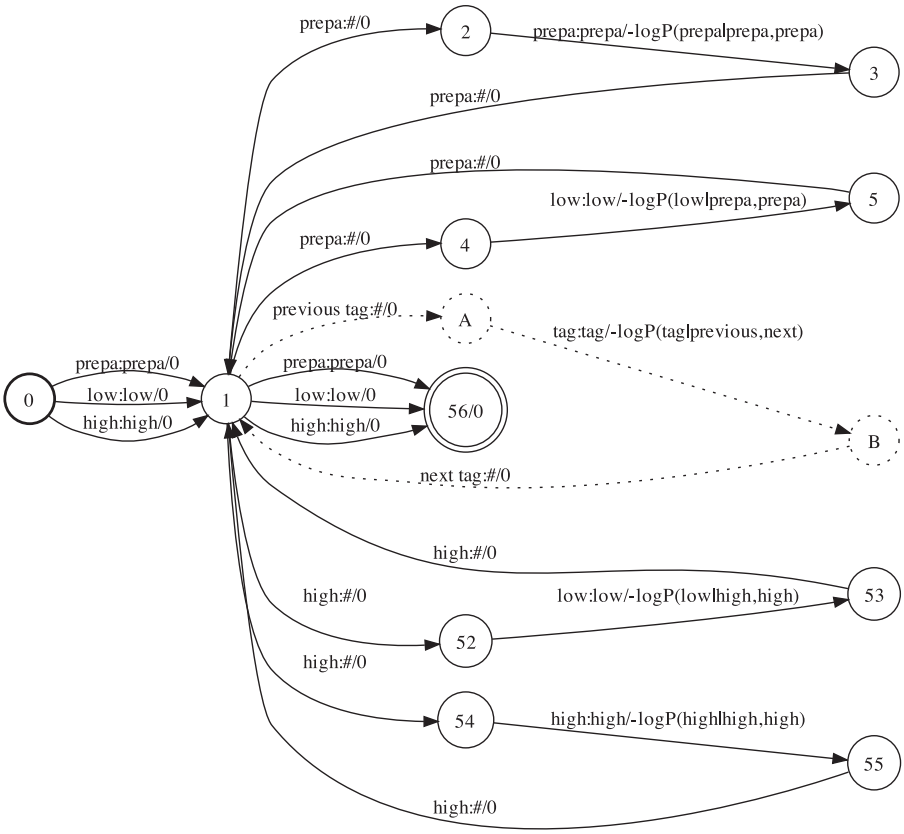

Fig. 3. Transducer *Mod* modeling all the contextual probabilities $P(s_i|s_{i-1}, s_{i+1})$.

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

3.4.2. Equation solving

In practice, to resolve Eq. (1), we projected the problem into the Finite-State Machine (FSM) paradigm by using weighted finite-state transducer, as presented in Mohri et al. (2002).

To do that, we have to represent the model containing the probabilities $P(s_i|s_{i-1}, s_{i+1})$ for all the 3-tuples (s_{i-1}, s_i, s_{i+1}) in a transducer representation. This FSM will be called M. Fig. 3 shows the topology used to represent all these probabilities using the FSM formalism. As we can see in this figure, input values of the transducer are used to represent 3-tuples (s_{i-1}, s_i, s_{i+1}), while output values correspond to the effective tag-sequence: the global cost of a hypothesis tag-sequence, according to the contextual tag model, is the cost of the path of which the output values correspond to this tag-sequence.

We have chosen to handle the costs by using the tropical semi-ring: the cost value of a 3-tuple (s_{i-1}, s_i, s_{i+1}) corresponds to the values of $-logP(s_i|s_{i-1}, s_{i+1})$. Notice that in this paper, *null* input or output values are represented by the symbol '#' (sharp).

To apply *Mod* to the search space representing all the hypotheses, we had to represent these hypotheses by using a FSM formalism compatible with *Mod*. This FSM will be called *Hyp*. The costs in *Hyp* are the confidence measure values $c(s_i)$ given by the *BoosTexter-like* classifier for each of the three possible tags (classes of spontaneity).

In order to reduce the number of paths in *Hyp*, we have taken into consideration the fact that some parts of paths can be factorized: as all the 3-tuples $(*, s_i, s_{i+1})$ will be followed by 3-tuples $(s_i, s_{i+1}, *)$, we have used this property during the generation of the topology of *Hyp*. Then, to apply the contextual tag-model represented by *Mod* to the hypotheses represented by *Hyp*, we make the composition of transducers: $Hyp \circ Mod$.

Last, as we use the tropical semi-ring to handle FSM costs to find the minimum cost path in $Hyp \circ Mod$, it means that an approximated Viterbi decoding is used: when multiple paths are identically labeled, the tropical semi-ring selects the minimum cost path only. The output values of the minimum cost path corresponds to the final tag-sequence hypothesis.

Due to the very small number of tags and to the small average number of speech segments detected in an audio file, this process is very fast: about 3 s of computation time with a standard PC to process the 11 files (11 h of speech/3,814 segments).

3.5. Experiments

The experimental corpus (as described in Section 2.1) consists of 11 audio files from radiophonic recordings.

Table 4
Performance of the ASR system according to speech category in terms of WER and NCE. The number of words, the number of speech segments and the duration according to each class of spontaneity are also included.

Class of spontaneity	Duration	# Segments	# Words	WER (%)	NCE
Prepared	3h40	1,228	39,984	10.1	0.358
Low spontaneous	3h50	1,339	44,245	18.4	0.315
High spontaneous	3h30	1,247	42,515	28.5	0.237
Total	11h00	3,814	126,744	15.0	0.331

For the experiments, we used the *Leave-one-out cross validation*: 10 files were used for training, 1 for the evaluation and this process is repeated until all files have been evaluated. Due to the lack of data, no development corpus was used to optimize the training process by avoiding overtraining. The training files will be used at the segment level for the classification process, and at the global level for the estimation of $P(s_i|s_{i-1}, s_{i+1})$.

3.5.1. ASR system performance

The acoustic, linguistic and ASR confidence measure features used as descriptors to characterize the speech spontaneity are extracted from the LIUM ASR system described in Section 3.2. Table 4 presents the results in terms of word error rate (WER) and normalized cross entropy (NCE) of the LIUM ASR system on the experimental data. The number and the duration of speech segments according to each spontaneity class are also shown. These manually labeled data were not included in the training or development corpus of the acoustic and linguistic models used in the ASR system.

Table 4 shows that the global performance of the ASR, with a WER of 15% and a NCE of 0.331, is very good for French Broadcast News processing. As expected, the more the speech is fluent, the more the WER is low: from 10.1% for speech segments manually annotated as *prepared* to 28.5% for *high spontaneous* speech segments. It is interesting to note the correlation between the WER obtained on ASR system transcriptions and the subjective annotation in speech spontaneity level. These results show the need of a particular focus on spontaneous speech detection.

3.5.2. Automatic categorization and detection of spontaneous speech

In order to measure the gain provided by the different kinds of descriptors, four conditions were evaluated:

- Linguistic features only on reference transcriptions: *ref:ling*.
- Linguistic features only on automatic transcriptions: *asr:ling*.
- Acoustic features only on automatic transcriptions: *asr:acou*.
- All features (acoustic, linguistic and ASR confidence measures) on automatic transcriptions: *asr:all*.

Thereby, we can compare results on reference and on automatic transcriptions provided by the ASR system. Note that we could not have results on reference acoustic features because obtaining scores for this type of information could not be manually done. Moreover, even if these features could be computed from manual transcriptions with an ASR system, there is still a bias due to the pronunciation dictionary, which does not necessarily integrate the exact pronunciation of each reference word.

We firstly focus on categorization and detection at the segment level. Table 5 presents the detection results (in terms of precision, recall and F-measure) for each spontaneity class. As we can see, the detection performance on the *low spontaneous* segments is low; this is not surprising as these segments can be easily misclassified as *prepared speech* on one side or *high spontaneous* on the other side.

The recall metric allows us to measure the coverage of our detection system while the precision metric gives information about its classification accuracy.

When focusing on the local decision (i.e. at segment level), we can see a drop between the performance achieved on the linguistic features using the reference transcriptions *local(ref:ling)* and using the automatic transcriptions *local(asr:ling)*. This fall is likely due to ASR transcription errors. However, we note that this gap is offset, and even improved, when we use the acoustic features *local(asr:acou)*, which are more robust to ASR transcription errors. Thus, the use of a classifier based on all the automatically extracted acoustic, linguistic and ASR confidence measure features *local(asr:all)* improves global performance. Indeed, we obtain better results whatever the class of spontaneity or the metric used (recall, precision or F-measure). Pitch estimation is included in the acoustic features. In our experiments, integrating pitch permits a very slight improvement, but not crucial in this work, increasing recall and precision with less than 0.2 points.

We then took into account the results obtained on the allocation of spontaneity classes for each segment in order to make a new decision: now, we use a global decision process. Two experiments were realized to measure the impact of this additional method:

- *global(ref:ling)*: the global model using labeling segment results obtained with features extracted from reference transcriptions *ref:ling*,
- *global(asr:all)*: the global model using labeling segment results obtained with features extracted from automatic transcriptions *asr:all*.

By examining the results of the *global(asr:all)* condition in Table 5, we observe that the probabilistic contextual model applied to the *local(asr:all)* condition allows the system to greatly improve the classification performance, whatever the spontaneity class or the metric used. Moreover, we note that performance is also better when using the *global(asr:all)* system in comparison with the *global*

Table 5
Precision, recall and F-measure of the speech segment classification according to 3 categories: *prepared speech*, *low spontaneous* and *high spontaneous*.

Features	Local decision				Global decision	
	ref:ling	asr:ling	asr:acou	asr:all	ref:ling	asr:all
Prepared speech						
Precision	56.0	53.0	56.3	**59.8**	61.6 (+9.1%)	**66.8** (+10.5%)
Recall	64.1	61.8	59.0	**65.1**	66.5 (+3.6%)	**69.6** (+6.5%)
F-measure	59.8	57.1	57.6	**62.3**	64.0 (+6.6%)	**68.2** (+8.7%)
Low spontaneous						
Precision	43.8	40.7	44.3	**47.0**	46.9 (+6.6%)	**54.3** (+13.4%)
Recall	37.7	31.7	41.2	**43.5**	42.8 (+11.9%)	**51.8** (+16.0%)
F-measure	40.5	35.6	42.7	**45.2**	44.8 (+9.6%)	**53.0** (+14.7%)
High spontaneous						
Precision	65.2	58.0	60.5	**66.7**	70.3 (+7.3%)	**73.0** (+8.6%)
Recall	65.9	60.4	62.2	**66.3**	71.5 (+7.8%)	**73.5** (+9.8%)
F-measure	65.4	59.2	61.3	**66.5**	70.9 (+7.8%)	**73.2** (+9.2%)

In bold the best classification results obtained.

(ref:ling) system. This could be explained by the fact that *global(asr:all)* takes into account specific ASR features (acoustic features and ASR confidence measures). If we compare each class of spontaneity, we can see that the new gains obtained with the *global(asr:all)* method are disparate. Indeed, the *low spontaneous* class obtains the best improvement, with a relative gain of 16.0% on recall and of 13.4% on precision. The better results are directly related to the intermediate position of this class, which is favorably influenced by the best performance of *prepared* and *high spontaneous* classes.

To understand how the segments were labeled and what is the precise impact of the global statistical method, Table 6 presents the confusion matrix of *local(asr:all)* (classification at the segment level with features extracted on automatic transcriptions). Then, Table 7 shows the confusion matrix computed on the *global(asr:all)* probabilistic model results. The confusion matrix represents the total number of speech segments automatically labeled with a class of spontaneity. Thereby, these results allow us to estimate the recall and precision for each class. Consider an example of the confusion matrix in Table 6, specifically the case of the *high spontaneous* speech class:

- Reading the table from left to right, we have to classify 1,228 segments as *prepared speech*. Of these 1,228 segments, 799 segments were actually correctly labeled as *prepared*, 343 were falsely labeled as *low spontaneous*, and 86 were falsely labeled as *high spontaneous*.
- Reading the table from top to bottom, 1,337 segments were automatically classified as *prepared speech*. Of these 1,337 segments, 799 have actually been correctly classified as *prepared*, 430 should have been categorized as *low spontaneous*, and 108 should have been categorized as *high spontaneous*.

By analyzing the confusion matrix in Table 6 for *local(asr:all)*, we see that the *low spontaneous* class is the actual weak point of our system. However, this does not seem surprising. Indeed, the initial goal of this speech spontaneity detection system is to retrieve as accurately as possible the segments containing *high spontaneous speech*. We note that the number of falsely labeled segments between the *prepared* and the *high spontaneous* classes is quite high. And this is this kind of errors that we would like to correct with our global method. If we focus on Table 7, we realize that taking into account the neighboring segments in the decision process can greatly reduce this type of errors, while improving results on the *prepared* class. Thus, we find that we can reduce by 38.1% the misclassification between *prepared* and *high spontaneous* classes.

A focus on the detection of *high spontaneous* speech segments is interesting because such segments are harder to automatically process. For example, for ASR processing, their detection is particularly important when applying specific solution as the one we proposed in Dufour et al. (2010b). By accepting all the classification propositions, our method achieves a 73.0% precision for high spontaneous speech detection with a 73.5% recall measure, as presented in Table 5. More precisely, 83.5% of high spontaneous detection errors are due to confusion between low and high spontaneous speech.

As we earlier mentioned in Eq. (1), we used the classification confidence measures $c(s_i)$ provided at the segment level by the *AdaBoost* classification tool for each spontaneity class. Moreover, with the combination of the scores $c(s_i)$ with the probabilities $P(s_i|s_{i-1}, s_{i+1})$ given by the contextual model, it is possible to filter the hypothesis class by applying a threshold on the computed score $c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$.

Fig. 4 presents the detection performance obtained by changing the threshold on classification score for *high spontaneous* segments. It shows the performance obtained by the classifier (*local*) when varying $c(s_i)$, and it also shows the results obtained using the *global* method.

We can see that our system becomes more accurate (precision increases) when we make less decisions (recall

Table 6
Confusion matrix on classification results obtained with the *local(asr:all)* method.

		Estimated class			
		Prepared	Low sponta.	High sponta.	Total
Correct class	*Prepared*	799	343	86	1,228
	Low spontaneous	430	582	327	1,339
	High spontaneous	108	312	827	1,247
	Total	1,337	1,237	1,240	3,814

Table 7
Confusion matrix on classification results obtained with the *global(asr:all)* method.

		Estimated class			
		Prepared	Low sponta.	High sponta.	Total
Correct class	*Prepared*	855	315	58	1,228
	Low spontaneous	363	694	282	1,339
	High spontaneous	62	268	917	1,247
	Total	1,280	1,277	1,257	3,814

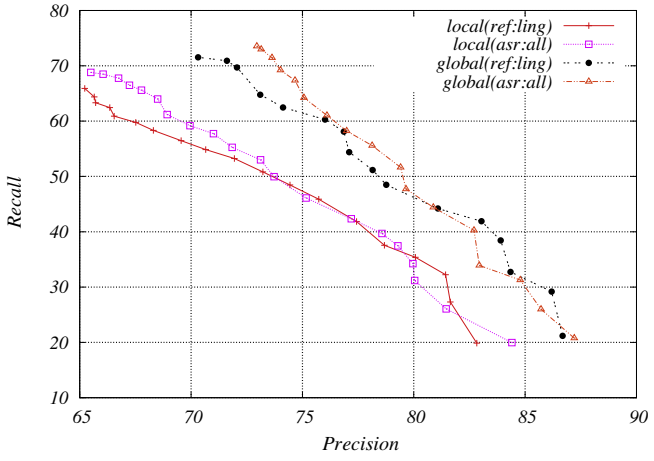

Fig. 4. Detection performance of high spontaneous segments according to a varying threshold on the classification score.

decreases). This possible thresholding adapts the use of the classification method: the best compromise between recall and precision could be done for the targeted application. Moreover, we could see that *ref:ling* and *asr:all* follow the same trend for both method levels (*local* and *global*), even when the decision threshold varies. We can then conclude that the acoustic features, linguistic features, and ASR confidence measures *asr:all* extracted using a transcription system achieve the same performance as the linguistic features extracted from the reference transcriptions *ref:ling*. The information loss due to transcription errors can be compensated by other features extracted from an ASR system.

4. Recognition of speaker roles using speech spontaneity features

4.1. Applying spontaneous speech detection to characterize speaker roles

In this part, the spontaneous speech detection system is applied to an annotated corpus in speaker roles, with two main objectives:

- First, we want to see if a class of spontaneity might be useful information to help to characterize the speaker role in an audio document. This first analysis follows some previous works realized on speaker role in Barzilay et al. (2000) and Liu (2006) or topic identification in Peskin et al. (1993) and McDonough et al. (1994). The main goal of this field is to extract features (acoustic/linguistic/etc.) in order to identify the structural information in broadcast news. For example, this identification could be useful for information retrieval in automatic transcriptions as presented in Amaral and Trancoso (2003). In addition, we continue the idea developed in Bigot et al. (2010) which highlights the fact that speaker role detection could help to enrich interaction sequences between speakers; this is useful in the field of automatic audiovisual content-based indexing and structuring.
- Second, this study may be useful to indirectly validate our detection system. Indeed, before starting experiments, we suspected that some speaker roles of audio data have a predominant spontaneity class. For example, we intuitively think that the *Interviewee* role usually belongs to the *high spontaneous* class of spontaneity, while the *Radio-program presenter* role would better fit in the *prepared* class.

The next section will give details about the EPAC project, and especially information about the annotated corpus in speaker roles of audio data. We then present the experiments made on this corpus and the results obtained.

4.2. The EPAC project

4.2.1. General presentation

The EPAC project, described in Estève et al. (2010), dealt with French unstructured audio data. The project started in March 2007 and ended up in August 2010. The

main concern of this project was to propose information extraction and document structuring methods to process audio data. All information channels are taken into account: signal segmentation (speech/ music/jingle/etc.), speaker identification, speech recognition, topic identification, emotion detection, etc. In particular, this project focused on automatic spontaneous speech processing.

The audio data processed during the EPAC project comes from the ESTER 1 campaign (Galliano et al., 2005) that compared performance of automatic speech recognition systems on French Broadcast News. The ESTER 1 corpus contains about 1,800 h of radio programs which come from three French radios: France Info, France Culture and RFI, broadcasted between 2003 and 2004. Only 100 h of the corpus were initially manually annotated.

The EPAC corpus is composed by manual annotations and transcriptions of 100 h of shows containing a major part of spontaneous speech (interviews, debates, and talk shows) and included in the initially untranscribed 1,700 h distributed during the ESTER 1 evaluation campaign.

4.2.2. Manual annotation of speaker roles

The EPAC corpus includes additional information on speakers and transcribed shows. More precisely, the role, the function and the profession of each speaker appearing in the corpus were manually specified when available. Thus, the same speaker only have one general role (e.g., *Guest*, *Interviewee*, *Commentator*, etc.) but could be refined with up to two other labels (e.g. for *Guest*: *politician/prime minister*), based on available information. In the case of a debate, a clarification regarding the views of speakers is added, as they are favorable or not to the question. All the EPAC corpus had been manually annotated in roles by one expert linguist. As the speech spontaneity annotation and the manual transcription, the manual speaker role labeling of the EPAC corpus is publicly available through the ELRA catalog (see Section 2.1).

Table 8 shows the proportion (in number of speakers, number of speaker turns, and duration) of the 10 manually labeled speaker roles of the EPAC corpus. Here we introduce the *speaker turn* term which could be defined by a change of the active speaker. A speaker turn may have several speech segments and is variable in terms of duration. We can see that the *Auditor*, *Expert*, *Guest* and *Interviewee* are the most represented speaker roles in terms of number of speakers. This seems normal since the corpus is oriented to broadcast news audio data. In this table, we have to note that the *Auditor* and *Interviewee* speaker roles appeared in a reduced number of segments compared to the number of speakers, which could be explained by their function: these speakers occasionally operate on shows, with short interventions.

Table 8
Proportion (in number of speakers, number of speaker turns and duration) of focused manually labeled speaker roles of the EPAC corpus.

Speaker role	# Speakers	# Speaker turns	Duration
Auditor	238	424	3h09
Commentator	135	182	4h19
Expert	151	1,527	16h23
Guest	134	2,813	26h46
Interviewee	116	438	4h02
Interviewer	31	227	0h30
Radio-program presenter	191	5,223	19h46
Reporter	11	18	0h10
Special correspondent	85	113	1h38
Other	45	220	1h47
Total	1,137	11,125	78h30

Each role is defined as follows:

- **Auditor**: radio listener which occasionally occurs during a program.
- **Commentator**: broadcaster or writer who reports and analyzes events in the news during a broadcast show.
- **Expert**: has knowledge in a particular field.
- **Guest**: comes to talk about the latest news.
- **Interviewee**: answers questions raised by interviewer or presenter.
- **Interviewer**: only leads an interview. In this broadcast news corpus, there are a few interviewer roles: the presenter role is generally preferred, as this speaker has multiple functions (see description below).
- **Radio-program presenter**: has multiple functions inside a show: can host a talk show, may take calls from listeners, or has the responsibility of giving news, weather information, etc.
- **Reporter**: only investigates and reports news stories.
- **Special correspondent**: particular journalist who contributes reports on particular subjects from a remote location.
- **Other**: all the other speaker roles which are not studied (lack of data).

4.3. Speaker role and speech spontaneity relationship: a qualitative study

4.3.1. Automatic detection of speech spontaneity

The EPAC corpus is used for our experiments. The first step of this study is to automatically detect the kind of speech of audio segments. Thereby, we have to transcribe the 80 h of the EPAC corpus that contains manually labeled annotations in speaker roles. The automatic speech transcription is made with the LIUM transcription system (see Section 3.2). Since the manual annotation of the speaker roles of audio data had been done on the reference segmentation, we decided to keep this segmentation for the decoding process. Indeed, without this segmentation, it would have been impossible to do a precise quantitative study.

We then extracted the acoustic, linguistic, and ASR confidence measure features for each speech segment in order to classify the segments according to the speech spontaneity. The training data consists of the 11 manually

annotated files with classes of spontaneity presented in Section 2.1. Table 9 presents the total duration and number of segments automatically labeled according to the three classes of spontaneity: *prepared*, *low spontaneous* and *high spontaneous*.

As we can see, the duration and the number of segments are homogeneously distributed between each class of spontaneity. In this study, we particularly focus on the first level of role annotation. Indeed, the next two levels are not very informative for our study because of their proportion being too small in the corpus (for example, the second level speaker role "humorist" only appeared in two segments). This is the reason why the total duration and the total number of segments in Table 8 is lower than the ones in Table 9, which represents the automatically annotated segments.

4.3.2. Speaker role characterization

We carried out the study on the annotated corpus in speaker roles. Firstly, we assigned each annotated segment with a class of spontaneity, as presented in Section 4.3.1. As for a same role some speakers only occur in a few number of segments while others got a lot of interventions, we normalized them in order to have the same weight for each speaker. Thus, we began by computing the average proportion (in terms of segments) of each class of spontaneity for each speaker. Then, we divided, for each role and each class of spontaneity, the sum of the average proportions of that class of spontaneity by the total number of speakers belonging to that role. Table 10 presents the proportion (in percentage) of manually labeled speaker roles for each class of spontaneity.

We can see that certain classes of spontaneity predominate in some speaker roles. Indeed, we notice that the *Commentator*, *Special correspondent*, and in a smaller proportion, the *Radio-program presenter* speaker roles contain a large proportion of *prepared* speech. These findings seem to be in line with the function of each role, as these speaker require certain preparation before the intervention. On the other hand, the *Guest*, *Auditor*, *Interviewee* speaker roles are mainly classified as *high spontaneous* speech. This makes sense since these speakers usually do not know their interventions in advance. Finally, the *Expert* and *Interviewer* speaker roles do not have a predominant class of spontaneity, which could be explain by the fact that speakers alternate between a prepared speech (subject known in advance) and more spontaneous reactions when dealing with unknown questions. But this uniform distribution of classes of spontaneity can be itself an interesting criterion to characterize these roles. It is not the lack of predominance of a particular class of spontaneity but the distribution of speech segments into these classes which is informative indeed, especially between the two extreme classes "*prepared vs. high spontaneous*" speech.

Table 9
Proportion (in duration and number of segments) of automatically classified speech segments according to the three classes of spontaneity.

Class of spontaneity	Duration	# Segments
Prepared	25h39	30,568
Low spontaneous	24h41	23,053
High spontaneous	28h50	29,554
Total	79h10	83,175

Results in Table 10 give the general tendency of each speaker role and its associated classes of spontaneity. But as we are also interested in comparing the speaker distribution inside its speaking role, Fig. 5 presents the distribution of speakers on the manually annotated speaker roles associated to these two classes of spontaneity. This figure does not present the results of the *Interviewer* and *Special correspondent* as the number of speakers or their total duration was not sufficient to make a clear study.

In order to better understand how to read Fig. 5, let's take a look at the *Commentator* role. We have the *blue* histograms which represent the distribution containing the *prepared* class, while the *green* ones represent the *high spontaneous* class of speech. As we can see, the first green histogram, with the *0–0.1 X* axis label, reaches the value of 0.9. It means that 90% of the speakers has a proportion of *prepared* speech between 0 and 10% during all their interventions: their speech is clearly not prepared for most of the time. Moreover, if we look at the *blue* histogram between 0.9 and 1, around 60% of their interventions are highly considered as *prepared* speech. Then, we can conclude that a majority of "Commentator" speakers prepare their interventions before speaking, which is consistent with previous findings Table 10. Globally, for each speaker role, we get similar results to the ones before: *Auditor*, *Commentator* and *Radio-program presenter* speakers have a tendency to speak more fluently than the *Guest* and *Interviewee* speakers. A confusion still exists between *prepared* and *high spontaneous* speech for the *Expert* speaker.

4.4. Detecting speaker roles with a speech spontaneity detection system

4.4.1. Related work

Even if detecting speaker roles is a recent research field, some work has focused on extracting this information from audio documents. Two information levels are generally

Table 10
Proportion (in percentage) of each class of spontaneity on manually labeled speaker roles.

Speaker role	Prepared	Low sponta.	High sponta.
Auditor	20.9	32.6	**46.5**
Commentator	**84.6**	12.3	3.1
Expert	32.2	34.4	33.4
Guest	26.7	22.9	**50.4**
Interviewee	25.1	32.3	**44.5**
Interviewer	26.3	36.9	36.8
Radio-program presenter	**56.6**	26.8	16.6
Special correspondent	**79.8**	14.0	6.2

In bold the best classification results obtained.

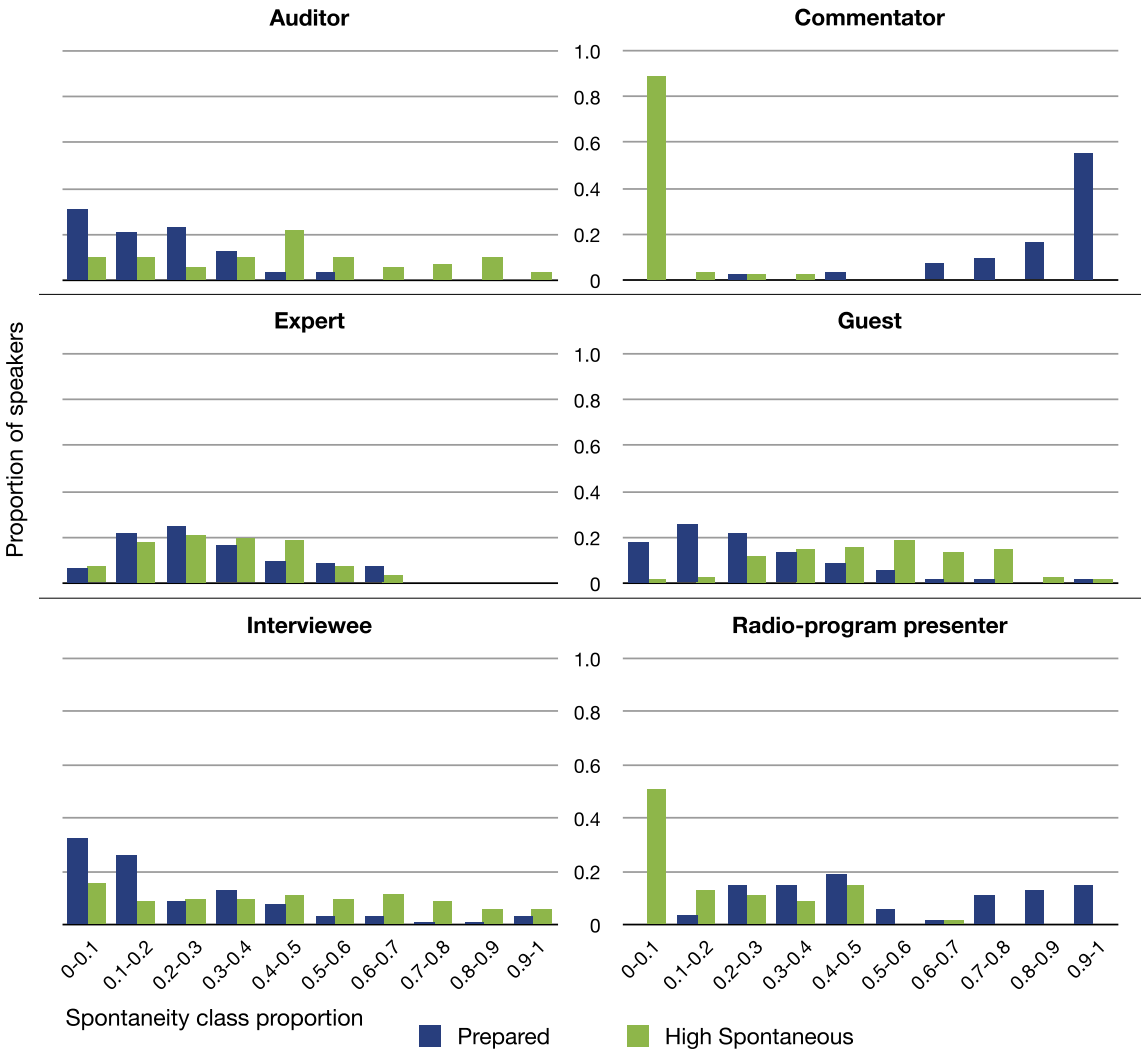

Fig. 5. Distribution of speakers on the manually annotated speaker roles associated to the "prepared" and "high spontaneous" classes of spontaneity.

used to identify speaker roles: acoustic/prosodic features (Salamin et al., 2009; Bigot et al., 2010) or lexical features (Barzilay et al., 2000; Garg et al., 2008; Liu, 2006; Damnati and Charlet, 2011). These related works usually propose an automatic classification process to assign a role to each speaker. In Barzilay et al. (2000), the authors propose to classify three speaker roles (*Anchor*, *Journalist* and *Guest*) using lexical (word n-grams) and duration features (segment duration) extracted from automatic speech transcriptions with a *boosting* algorithm and a maximum entropy modeling. Moreover, the authors use the surrounding information as a feature in the decision process. With a manually labeled speaker turn but with an automatic transcription, the method achieved a correct classification of 80% of the speech segments. A similar approach had been recently developed by Damnati and Charlet (2011), where the authors try to assign three speaker roles in TV broadcast news programs. Experiments are made on a fully automated system with an automatic segmentation, speaker diarization, and transcription. This method allows the authors to correctly classify 86% of speaker turns. Authors in Liu (2006) investigate the same speaker roles and use of HMM and maximum entropy model on manual transcriptions and manually labeled speaker turns and roles. Classification with an accuracy of 80% is obtained with the combination of both models. In Bigot et al. (2010), the authors use acoustic, temporal and prosodic features for classifying the same five speaker roles (*Anchor*, *Journalist*, *Other*, *Punctual Journalist* and *Punctual Other*) in documents containing mainly spontaneous speech. These extracted features associated with a supervised classification algorithm allows the system to achieve a 92% of good speaker role attribution.

Moreover, in Vinciarelli (2007), the authors introduce the concept of *Social Network Analysis* (SNA) for speaker role recognition. The idea behind SNA is that a specific speaker role (*actors*) interact with others during *events*. These interactions may find the role associated to each involved speaker. In Vinciarelli (2007), SNA is combined with approaches using a speaker duration analysis. This

combination achieves an 85% accuracy of speaker role classification of the recording time (6 roles and 11 speakers). This work has been continued in Salamin et al. (2009). Moreover, SNA had also been successfully applied in Garg et al. (2008), in which the authors combined SNA with a classification approach based on lexical information (word n-grams) using *AdaBoost* algorithm. A 70% accuracy of speaker role classification of the recording time is achieved with this approach.

We note that in the literature, most of these studies try to recognize from three (Barzilay et al., 2000; Damnati and Charlet, 2011) to six (Vinciarelli, 2007; Salamin et al., 2009) speaker roles.

4.4.2. General approach

In order to detect speaker roles in broadcast news audio data, we propose to use an existing method already successfully applied to spontaneous speech detection. As seen in the preliminary study in Section 4, the speech spontaneity could be a useful feature for speaker role recognition methods. Moreover, the previously presented approaches for speaker role identification generally use acoustic or linguistic features. These two information sources are already used by our speech spontaneity detection system as seen in Section 3. Speech spontaneity detection and speaker role classification can both be seen as a multiclass classification problem which could be solved with a classification algorithm (Support Vector Machine, AdaBoost, etc.).

By building on these similarities and on conclusions made by our preliminary study showing the link between speech spontaneity and speaker roles (see Section 4), we propose to directly apply the two-step speech spontaneity detection method (*Local* and *Global* decision). Instead of using speech spontaneity information inside a speaker role recognition system, we will classify speaker roles with a speech spontaneity system.

No modification had been made on the detection method: the same acoustic and linguistic features as those presented in Dufour et al. (2009b,a, 2010b) are automatically extracted with a transcription system, and a contextual model is used to take into consideration neighboring context. The only difference is the size of the classified data used as input: the segment level is not studied anymore, we now work at the speaker (all interventions of a speaker in the audio document) and speaker turn (change of the active speaker) levels. Indeed, this does not change anything in the method process. For the local detection, each speaker is classified, which generally covers many segments and sometimes many speaker turns. In addition, the contextual decision model uses information about surrounding speaker turns (the same speaker may have many speaker turns inside an audio document). During the global process, various roles may be associated to a same speaker: in that case, the most frequent one, in terms of speaker turns, is chosen (one speaker could only have one general associated role in this corpus, as exposed in Section 4.2.2). Finally, instead of trying to assign one of the 3 classes of spontaneity at each segment, each speaker will be associated with one of the 10 speaker roles defined in Section 4.2.2.

In the following section, we present the experiments made with this approach on the EPAC corpus. To evaluate the speaker role detection method, we will firstly classify speaker roles in particular and favorable conditions: an automatic transcription is used, but segmentation and speaker diarization remain manual. Then we will propose to use this method in a fully-automated way: we have to deal with problems of automatic speech transcription which could influence speaker role recognition (i.e. automatic segmentation and speaker diarization).

4.4.3. Semi-automated speaker role detection system

The EPAC corpus, composed of 121 audio files (see Section 4.2), has been used for our experiments. To evaluate the performance of our speaker role detection system, we followed the same *Leave-one-out* cross validation, already presented in Section 3.5. Manual segmentation and manual speaker diarization are used (we know exactly who speaks and when). Results presented in this study will only evaluate the speaker role identification process: problems of automatic segmentation and diarization are not addressed. The experiments made with this semi-automated system had already been presented in Dufour et al. (2011).

Finally, following the approach previously presented, we automatically associated a role to each speaker using the speech spontaneity detection method. Table 11 presents the speaker role recognition results obtained on the EPAC corpus in terms of precision, recall, and F-measure on each step of the decision method: *Local* then *Global* decision.

As we can see, classification performance may differ depending on the speaker role. Indeed, if we focus on the *Local* process, i.e. the decision made at the speaker level, we can note that *Auditor* and *Presenter* recognition performance is already very high, and that the *Experts* obtain acceptable results. But on the other hand, identification performance of *Reporters* and *Others* is very low. This is not surprising as the first one is very infrequent in terms of number of speakers and duration (see Table 8), and the second one regroups a large variety of different speaker roles. Finally, we obtain interesting preliminary results for the *Commentators*, *Correspondents*, *Guests*, *Interviewees* and *Interviewers*. Although performance in these classes is not yet satisfactory, we suspected that the contextual model would help to better categorize them with the help of classification results of surrounding speaker turns. The overall classification precision reaches 71.2% for this *Local* decision.

The contextual model applied to the classification results obtained at the *Local* decision (*Global*) permits the system to drastically improve recognition performance for *Commentator*, *Expert*, *Interviewee*, *Interviewer* and *Presenter*. Although performance is much better with the use of the *Global* process, it still remains weak for the *Other* speaker role. But this role is particular, since its number of

Table 11
Recall, precision and F-measure of the speaker role recognition process using the two-step speech spontaneity detection method (Local then Global decision).

Speaker role	Local			Global		
	Recall	Prec.	F-mea.	Recall	Prec.	F-mea.
Auditor	92.4	92.8	92.6	91.6	**95.6**	**93.6**
Commentator	60.7	57.8	59.2	**63.7**	**59.3**	**61.4**
Expert	73.5	71.2	72.3	**82.1**	**72.1**	**76.8**
Guest	61.2	66.1	63.6	**69.4**	65.0	**67.1**
Interviewee	51.7	45.5	48.4	**56.0**	**52.9**	**54.4**
Interviewer	61.3	57.6	59.4	**64.5**	**95.2**	**76.9**
Radio-program presenter	93.2	90.8	92.0	**96.3**	**92.0**	**94.1**
Reporter	18.2	50.0	26.7	9.1	33.0	14.3
Special correspondent	56.5	53.3	54.9	52.9	**56.3**	54.5
Other	17.8	34.8	23.6	**22.2**	**45.5**	**29.9**

In bold the best classification results obtained.

annotated data is really low compared to others. The *Auditor* speaker role recognition, which already reached high performance during the *Local* process, is much more accurate with the *Global* decision, with a slight decrease in recall. No improvement has been noted on classification performance of *Correspondents*. Finally, the *Reporters* classification performance decreases, still due to lack of data. This is the only class that see its F-measure decrease.

We now reach an overall 74.4% classification precision with the use of the entire detection method. Even if the conditions were favorable in these experiments, results seem satisfying considering the high number of speaker roles studied (10 roles), and considering some other speaker role recognition performance related in Section 4.4.1 and summarized in Bigot et al. (2010). A fully automated speaker role recognition system is now proposed.

4.4.4. Fully automated speaker role detection system

The previous experiments that we performed on speaker role recognition did not take into account problems due to speech segmentation and speaker diarization. Indeed, to propose a speaker role detection system in real conditions, an automatic segmentation and diarization system should be used. For these experiments, the LIUM segmentation and diarization tool, previously presented in Section 3.2, has been applied to the 121 audio files of the EPAC corpus. Obviously, we used the same data and the same process as the ones used in the previous section for building the semi-automated system. Finally, no change has been made on the speech spontaneity detection system.

A new difficulty induced by the use of an automatic process is the question of the evaluation metric. Indeed, it could not be the same metric as the one of the semi-automated system which just check that a role is well attributed to a speaker. To take into account problems of segmentation and speaker diarization, we will now evaluate our system in terms of duration. Thus, the evaluation process will check every 10 ms (at frame level) that the correct role had been detected. The evaluation metric is close to the one used, for example, in named entity detection (Jousse et al., 2009):

- Correct (C): the detection system correctly attributed the speaker role.
- Substitution (S): the detection system falsely attributed the speaker role.
- Deletion (D): no speaker role was attributed even though a speaker role exists in the reference.
- Insertion (I): a speaker role was attributed even tough no speaker role is identified in the reference.

The precision and recall metrics will now be computed as follows:

$$\text{Recall}_i = \frac{C_i}{C_i + S_i + D_i} \qquad \text{Precision}_i = \frac{C_i}{C_i + S_i + I_i} \tag{2}$$

Table 12 presents the speaker role recognition results obtained with the fully-automated system based on the spontaneous speech detection method. Results are evaluated at the duration level in terms of precision, recall, and F-measure.

Firstly, we can see that most of the speaker role classes follow the same tendency as the one observed in Table 11 with the semi-automated system. Indeed, the *Auditor* and *Radio-program presenter* speaker roles still obtain the highest recognition performance while the under-represented classes (*Reporter* and *Other*) still get poor detection results. However, we can note that the *Guest* speaker role gets highest results with this fully-automated system: it can be explained by the fact that the evaluation is now computed at duration level, as *Guest* is the most representative class at this level (see Table 8).

The main difference with this fully automated system is seen at the global decision performance. Indeed, it does not provide improvements as significant as those obtained with the semi-automated system. The automatic segmentation and diarization process could explain this difference: since the global method uses information of surrounding speaker turns, an error in the automatic segmentation can cause cascading errors in the whole document, which was not the case in the local process.

The fully-automated system, using the two-step speech spontaneity detection method, finally reaches a 76.7%

Table 12
Recall, precision and F-measure in terms of duration of the fully-automated speaker role recognition process using the two-step spontaneous speech detection method (Local then Global decision).

Speaker role	Local			Global		
	Recall	Prec.	F-mea.	Recall	Prec.	F-mea.
Auditor	79.8	81.8	80.8	72.5	**88.8**	**79.9**
Commentator	62.6	65.2	63.9	60.8	**71.4**	**65.5**
Expert	76.8	68.2	72.2	**81.1**	**70.0**	**75.1**
Guest	82.0	85.4	83.7	79.5	**86.3**	**82.8**
Interviewee	46.2	41.5	43.7	**47.5**	**46.5**	**47.0**
Interviewer	18.1	29.5	22.4	15.4	26.2	19.4
Radio-program presenter	87.2	88.7	87.9	**91.1**	86.7	**88.8**
Reporter	11.1	14.8	12.7	4.6	6.8	5.5
Special correspondent	55.7	45.0	49.8	46.2	**52.1**	49.0
Other	1.0	3.6	1.5	**1.6**	**3.9**	**2.3**
Total	76.3	76.5	76.4	**76.7**	**76.9**	**76.8**

In bold the best classification results obtained.

recall score and a precision of 76.9%. Although specialized speaker role detection system can reach more than 80% of good classification score, performance of the proposed system is very encouraging. Indeed, as already discussed, the number of speaker roles to classify is much higher in this study than those in the literature (maximum of 6 speaker roles). Moreover, this system was not initially designed to categorize speaker roles, this method can be enriched and improved with specialized speaker role features.

5. Conclusion

We firstly proposed an analysis of various acoustic and linguistic features extracted from an automatic speech recognition processing in order to characterize and detect spontaneous speech segments from large audio databases. To better define this notion of spontaneous speech, speech segments of an 11-hour corpus (French Broadcast News) had been manually labeled according to levels of spontaneity. This manual labeling helped to define three classes of spontaneity: *prepared*, *low spontaneous* and *high spontaneous* speech.

A two-step process had then been proposed to detect spontaneous speech segments. The first step is a classification process which consists in combining acoustic and linguistic features extracted from an ASR system. This classification method, performed at the segment level, allowed the system to associate a class of spontaneity to each speech segment. In the second step, we extended the classification process by using a probabilistic contextual tag-sequence model that takes into consideration information of surrounding segments: the classification becomes a global process. This method improved the results: 73.0% precision in the detection of *high spontaneous* speech segments, with a 73.5% recall measure, and a 66.8% precision and a 69.6% recall on *prepared* speech segments.

After applying a threshold on the scores obtained during the classification process, we demonstrated that the *high spontaneous* speech detection precision can reach more than 85%, but with a recall of 25%. This possibility of defining a threshold can adapt the use of this classification method by finding the best compromise between recall and precision for the targeted application.

This spontaneous speech detection provides a very useful piece of information which can be used by various applications, such as content-based indexing and structuring. In order to investigate this possibility, we analyzed the association of spontaneity classes with manually annotated speaker roles on audio broadcast news data. First, we applied our automatic speech spontaneity detection system on a manually annotated corpus, and made a qualitative study between the level of spontaneity and the speaker roles. Results obtained support our hypothesis: there is a correlation between the level of spontaneity and the speaker roles. Some speaker roles have a predominant class of spontaneity (for example, the *prepared* class for the *Commentator* role). Furthermore, more than the predominance of a particular class of spontaneity, it is the distribution of speech segments into these classes which is informative: this information should be exploited to define the speaker role of audio data.

Following this study, we proposed an original approach to detect speaker roles using our automatic spontaneous speech detection system. Experiments made on the EPAC corpus (80 h and 10 speaker roles) shown that features and approaches initially designed to detect speech spontaneity in audio documents could directly be applied to classify speaker roles. The proposed method allows the system to assign the correct role to 74.4% of the speakers with the semi-automated system (automatic transcription but manual segmentation and speaker diarization) and reaches a correct labeling 76.8% of the duration with the fully-automated system.

References

Amaral, R., Trancoso, I., 2003. Segmentation and indexation of broadcast news. In: ISCA Workshop on Multilingual Spoken Document Retrieval (MSDR), Hong Kong, China, pp. 31–36.

Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S., 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In: Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI), pp. 679–684.

Bazillon, T., Estève, Y., Luzzati, D., 2008. Manual vs assisted transcription of prepared and spontaneous speech. In: The sixth international conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morroco, pp. 1067–1071.

Bigot, B., Ferrané, I., Pinquier, J., André-Obrecht, R., 2010. Speaker role recognition to help spontaneous conversational speech detection. In: International workshop on Searching Spontaneous Conversational Speech (SSCS), Firenze, Italy, pp. 5–10.

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

Caelen-Haumont, G., 2002. Perlocutory values and functions of melisms in spontaneous dialogue. In: Proceedings of the First International Conference on Speech Prosody (SP), pp. 195–198.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Damnati, G., Charlet, D., 2011. Robust speaker turn role labeling of tv broadcast news shows. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic.

Deléglise, P., Estève, Y., Meignier, S., Merlin, T., 2009. Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? In: Conference of the International Speech Communication Association (INTERSPEECH), Brighton, United-Kingdom, pp. 2123–2126.

Duez, D., 1982. Salient pauses and non salient pauses in three speech style. In: Language and Speech, vol. 25, pp. 11–28.

Dufour, R., Estève, Y., Deléglise, P., Béchet, F., 2009a. Local and global models for spontaneous speech segment detection and characterization. In: Automatic Speech Recognition and Understanding (ASRU), Merano, Italy.

Dufour, R., Jousse, V., Estève, Y., Béchet, F., Linarès, G., 2009b. Spontaneous speech characterization and detection in large audio database. In: 13th International Conference on Speech and Computer (SPECOM), St. Petersburg, Russia.

Dufour, R., Bougares, F., Estève, Y., Deléglise, P., 2010a. Unsupervised model adaptation on targeted speech segments for LVCSR system combination. In: Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, Japan.

Dufour, R., Estève, Y., Deléglise, P., Béchet, F., 2010b. Automatic indexing of speech segments with spontaneity levels on large audio database. In: ACM Workshop on Searching Spontaneous Conversational Speech (SSCS), Firenze, Italy.

Dufour, R., Estève, Y., Deléglise, P., 2011. Investigation of Spontaneous Speech Characterization Applied to Speaker Role Recognition. In: Conference of the International Speech Communication Association (INTERSPEECH), Firenze, Italy.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., Farinas, J., 2010. The EPAC corpus: manual and automatic annotations of conversational speech in French Broadcast News. In: Language Resources and Evaluation (LREC), Valletta, Malta, pp. 1686–1689.

Eugenio, B.D., Glass, M., 2004. The kappa statistic: a second look. Computational Linguistics 30 (1), 95–101.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J., Gravier, G., 2005. The ESTER phase II evaluation campaign for the rich transcription of French Broadcast News. In: Conference of the International Speech Communication Association (INTERSPEECH 2005), Lisbon, Portugal.

Garg, P.N., Favre, S., Salamin, H., Hakkani-Tür, D., Vinciarelli, A., 2008. Role recognition for meeting participants: an approach based on lexical information and social network analysis. In: ACM Multimedia Conference (MM'08), Vancouver, Canada, pp. 693–696.

Goto, M., Itou, K., Hayamizu, S.A., 1999. A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In: Sixth European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, pp. 227–230.

Gravier, G., Adda, G., Paulson, N., Carr, M., Giraudel, A., Galibert, O., 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In: International Conference on Language Resources, Evaluation and Corpora (LREC), Istanbul, Turkey.

Hakkani-Tür, D., Tür, G., 2007. Statistical Sentence Extraction for Information Distillation. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, Honolulu, Hawaii, USA, pp. 1–4.

Heeman, P.A., Loken-Kim, K.-h., Allen, J.F., 1996. Combining the Detection and Correction of Speech Repairs. In: International Conference on Spoken Language Processing (ICSLP), vol. 1, Philadelphia, USA, pp. 362–365.

Jousse, V., Meignier, S., Estève, Y., Jacquin, C., 2009. Automatic named identification of speakers using diarization and ASR systems. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 4557–4560.

Lease, M., Johnson, M., Charniak, E., 2006. Recognizing disfluencies in conversational speech. IEEE Transactions on Audio, Speech and Language Processing 14 (5), 1566–1573.

Liu, Y., 2006. Initial study on automatic identification of speaker role in broadcast news speech. In: Human Language Technology Conference of the NAACL, NY, USA, pp. 81–84.

Liu, Y., Shriberg, E., Stolcke, A., Harper, M., 2005. Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection. In: Conference of the International Speech Communication Association (INTERSPEECH), Lisbon, Portugal, pp. 3313–3316.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M., 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on Audio, Speech and Language Processing 14 (5), 1526–1540.

Luzzati, D., 2004. Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané. In: Workshop Modélisation pour l'Identification des Langues (MIDL), Paris, France, pp. 13–17.

McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., Rohlicek, J., 1994. Approaches to topic identification on the switchboard corpus. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, Adelaide, Australia, pp. 385–388.

Meignier, S., Merlin, T., 2010. LIUMSpkDiarization: an open source toolkit for diarization. In: CMU Sphinx Users and Developers Workshop, Dallas, USA.

Mohri, M., Pereira, F., Riley, M., 2002. Weighted finite-state transducers in speech recognition. Computer Speech and Language 16 (1), 69–88.

O'Shaughnessy, D., 1993. Analysis and automatic recognition of false starts in spontaneous speech. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, Minneapolis, USA, pp. 724–727.

Peskin, B., Gillick, L., Ito, Y., Lowe, S., Roth, R., Scattone, F., Baker, J., Baker, J., Bridle, J., Hunt, M., Orloff, J., 1993. Topic and speaker identification via large vocabulary continuous speech recognition. In: Human Language Technology (HLT), Plainsboro, USA, pp. 119–124.

Rousseau, A., Bougares, F., Delglise, P., Schwenk, H., Estve, Y., 2011. LIUMs systems for the IWSLT 2011 speech translation tasks. In: Proceedings of IWSLT'11, San Francisco, CA, USA.

Salamin, H., Favre, S., Vinciarelli, A., 2009. Automatic role recognition in multiparty recordings: using social affiliation networks for feature extraction. In: IEEE Transactions on Multimedia, vol. 11, pp. 1373–1380.

Schapire, R.E., Singer, Y., 2000. BoosTexter: a boosting-based system for text categorization. Machine Learning 39, 135–168.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., 2012. The interspeech 2012 speaker trait Challenge. In:

Conference of the International Speech Communication Association (INTERSPEECH), Portland, OR, USA.

Shriberg, E., 1999. Phonetic consequences of speech disfluency. In: Proceedings of the International Congress of Phonetic Sciences (ICPhS), San Francisco, USA, pp. 619–622.

Siu, M., Ostendorf, M., 1996. Modeling disfluencies in conversational speech. In: International Conference on Spoken Language Processing (ICSLP), vol. 1. Philadelphia, USA, pp. 386–389.

Vinciarelli, A., 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Transaction on Multimedia 9 (6), 1215–1226.

Yeh, J.-F., Wu, C.-H., 2006. Edit disfluencies detection and correction using a cleanup language model and an alignment model. IEEE Transactions on Audio, Speech and Language Processing 14 (5), 1574–1583.