ON THE USE OF LINGUISTIC INFORMATION FOR BROADCAST NEWS SPEAKER TRACKING

William Antoni, Corinne Fredouille, Jean-François Bonastre

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{william.antoni,corinne.fredouille,jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

In this paper, we have explored a speaker characterization at two different linguistic levels, the lexical content and the syntactical form, in the context of Broadcast News (BN) speaker tracking task. The modeling of the information is done classically, by a n-gram approach applied on the word sequence for the linguistic content and on the syntactical tags issued from this sequence for the syntactical level (n-class modeling). The experiments were done on a subset of the BN rich transcription French evaluation campaign, ESTER. We have observed that lexical information is not useful for BN data while the syntactical information seems promising as it allows significant speaker identification and verification performance (until 40% of correct identification rate and 35% of EER).

1. INTRODUCTION

Information extraction in order to index multimedia recordings is an active research domain, largely supported by the different evaluation campaigns proposed for this field (Rich Transcription [1] or TREC [2] campaigns and the French ESTER [3] campaign). Different types of information may be extracted from sound signals: sound event, speaker information (turns, identity, gender), transcription, linguistic information, etc; each being related to a specific task. For example in Broadcast News, a request may be to retrieve the speech segments pronounced by a known speaker, referred to as the speaker tracking task. Indeed, this specific task consists in detecting portions of the document that have been uttered by a given speaker known beforehand and for which training data are available before the test/tracking stage. Two steps are classically performed by a speaker tracking system [4]: a speaker segmentation/clustering phase, which consists in providing speaker homogeneous segments along the radio Broadcast News shows, and a speaker detection process applied on each separate segment or on the different clusters (all the segments of a cluster are supposed to belong to one speaker only) in order to attribute the segments to the tracked speakers.
This paper is focused on the second step, the speaker detection process, which has also involved large research efforts over the last past decades. This interest has been mainly supported by the yearly NIST Speaker Recognition Evaluation campaigns. In this field, since the introduction of an extended data task, drastically increasing the amount of data, the community tends to model other types of information than acoustic information, like prosodic, phonetics [5][6], idiolectal [7], usually referred to as high level features [8]. The main goal of this paper is to investigate the use of syntactical information as a novel dimension able to characterize speakers. The basic idea of this approach is to take advantage of the structure of the speakers' discourse rather than its lexical content. Indeed, the assumption behind this idea is that syntactical information may be particularly interesting to track speakers in the context of Broadcast News shows, especially whether in the case of very short appearances.
Section 2 describes the method proposed to deal with syntactical information for the speaker recognition tasks (both identification and verification tasks). In section 3, preliminary studies are presented, which aim at demonstrating the interest of syntactical information compared to lexical one. Some conclusions following by some future works are given in section 4.

2. PROPOSED METHOD

In the literature, two main families of methods are proposed to characterize a language: the grammar-based approaches, describing the various structures accepted by a given language, and the stochastic model-based approaches, assigning a probability to various word sequences in a given language.
In this paper, stochastic models, largely used in the automatic speech recognition systems, are especially enlightened. However, even though they will be used for their capabilities to characterize the language, the goal of this study is to demonstrate that they can also be used to characterize speakers.

2.1. Modeling linguistic information with N-gram

Two main kinds of stochastic models is identified in the literature:

- lexical n-gram model (usually called "n-gram")
  - modeling of lexical inputs
  - based on the probability of a word knowing its history (n-1 preceding words)
  - depending on the language characteristic terms
  - characterize "what is said" but not "how it is said"
- syntactical n-gram model (usually called "n-class")
  - modeling of syntactical classes
  - based on the probability of a class knowing its history (n-1 preceding classes)
  - depending on the language characteristic structures
  - characterize not "what is said" but "how it is said"

If lexical n-gram models have already been studied for speaker characterization on conversational speech data, demonstrating its poten-

tiality [7] in this field, no study seems to have been carried out on syntactical aspect of the language, nor on BN data. However, the latter seems to be as relevant as the former one for the following reasons. First of all, syntactical n-gram models tend to be less dependent on the targeted application (or data processed) since they do not depend on the content of the discourse, but rather on the type of discourse (spontaneous speech, prompted speech, prepared speech, ...). In the context of tracking speakers in Broadcast News shows for instance, the syntactical constraint is less strong than lexical one. Indeed, for a tracked news anchor, the topics of talks may changed from a show to another one whereas the type of discourse may remain unchanged. Secondly, syntactical n-gram models should require less amount of training data than the lexical ones. This difference may be particularly relevant in applications with limited speech resource per speaker. This paper investigates this novel way of characterizing speakers. Both the lexical and syntactical modelings will be studied and compared to demonstrate the presence of specific information at the syntactical level able to characterize speakers.

2.2. Perplexity for similarity measure

Considering n-gram models, the probability of a sequence *W* of *k* words (or classes) is given by the following formula:

$$P(W_1^k) = P(w_1)\Pi_{i=2}^{k} P(w_i|w_1, ..., w_{i-1}) \quad (1)$$

involving the probability of each word (or class) *w* and the probability of the preceding word (or class) sequence, called history and noted *h* in the rest of the paper. Classically the perplexity criterion is used as similarity measure between a test and a n-gram model. Perplexity criterion is given by the next formulas.
Given the probability of a word sequence *W*, *P(W)*:

$$P(W) = \Pi_{i=1}^{k} P(w_i|w_1, ..., w_{i-1}) \quad (2)$$

the entropy of a word sequence *W*, noted *H(W)*, is defined by:

$$H(W) = -\frac{1}{k} log_2 P(W) \quad (3)$$

and the perplexity of this word sequence *W*, noted *PP(W)*, by:

$$PP(W) = 2^{H(W)} \quad (4)$$

The lower the perplexity of a sequence of words is, compared to model, the more it has a high probability of belonging to this model.

3. PRELIMINARY STUDY

The relevancy of syntactical information as opposed to lexical information to characterize speakers has been studied in the framework of the ESTER French evaluation campaign. Two main tasks, related to the speaker recognition field have been targeted: the automatic speaker identification and verification, considered as subtasks of speaker tracking, the targeted task in this paper.

3.1. ESTER campaign and corpora

The ESTER campaign has focused on the evaluation of rich transcription and indexing of radio broadcast news in French. It has been organized jointly by the Francophone Speech Communication Association (AFCP), the French Defense expertise and test center for speech and language processing (DGA/CEP), and the Evaluation and Language resources Distribution Agency (ELDA). It is part of the EVALDA project, sponsored by the Technolangue program, initiated by the French Ministry of Research and dedicated to the evaluation of language technologies for the French language. The ESTER campaign proposes different tasks, which can be classified in three main categories: orthographic transcription, event detection and speaker tracking, and data extraction. To respond to these different tasks, two main corpora have been available for participants:

- a train corpus of about 90 hours of manually transcribed radio broadcast news shows (including speech transcription, speaker identities and turns, name entity identification, ...); Transcribed data were recorded in 1998, 2000 and 2003, from four different sources. (1600 hours of not transcribed data are also available, but not used in this paper). This corpus will be named *ESTER-Train* in this paper.
- a test corpus of about 10 hours of transcribed radio broadcast news shows recorded from October to December 2004, from the same five sources of the training corpus plus an unknown source. This corpus will be named *ESTER-Test* in this paper.

In this preliminary work, both the reference transcriptions and segmentations are used. We focus on a set of speakers issued from the ESTER corpora. The speaker data are extracted and gathered from the *ESTER-Train* and *ESTER-Test* corpora, by selecting those having a sufficient quantity of data in train and test: the *n* speakers having at least *A* words as quantity in train and at least *T* words as quantity in test. Thus, for each of these *n* speakers, a *A*-size model is trained on the first *A* word segment of his overall train data (speakers may have more than *A* words on the *ESTER-Train*). The set of *n* speakers is noted S_n.

3.2. Targeted Tasks

The identification task consists in identifying a person among a set of people known by the system. It is performed here on the closed set of speakers S_n. The similarities between test and models are measured thanks to the perplexity criterion. If the lowest perplexity is reached by the model of speaker having producing the test segment, a correct identification is obtained. Finally, the rate of correct identification is used as performance measure.
The speaker verification task is the binary process of accepting or rejecting the identity claimed by a speaker under test. The client/target speaker set is the set of speakers S_n. Some additional impostor speakers are extracted from the *ESTER-Test* corpus. The constraint in terms of word quantity (T word segments) is also applied here for the test impostor selection. Like previously, similarities are measured thanks to a function of the perplexity criterion. However, as speaker verification scores should be normalized before taking the decision, a cohort based normalization is computed, using all the N (target) speaker models. The normalized similarity LPP_m^{norm} between a test and the model of speaker m is expressed as follows:

$$LPP_m^{norm} = (LPP_m - \frac{\Sigma_{i=1}^{N} LPP_i}{N}) \quad (5)$$

$$LPP_s = -log_2(PP(test|model_s)) \quad (6)$$

where LPP_s is the similarity between a test and a speaker model s.

3.3. System specification and Tools

In order to work at a syntactical level and train n-class model, we need a tagger, able to replace every word in transcription data by its syntactical class. The LIA-Tagger has been developed by the LIA and defines a set of 103 syntactical classes [9]. The perplexity criterion (defined in the previous section) is involved in this paper for the test and model comparison. The N-gram model training as well as the perplexity computation are performed thanks to the SRILM [10], which is a toolkit for building and applying statistical language models. It has been under development in the SRI Speech Technology and Research Laboratory since 1995. For this study, the following parameters are used for the n-gram modeling:

- the order of the n-gram model is 3
- the discounting method used for the back-off estimate of the 3-grams, is Witten-Bell, depending on the context of words and recommended for small size corpora.
- the minimal count for a n-gram (cut-off) to be included in the model is 1 (all of the n-grams met in train data are preserved).

3.4. Protocols

Two protocols are used for the experiments conducted in this study. For both protocols, the same amount of training data is used for each speaker issued from the speaker set S_n, leading to a A fixed to 1000 words in order to target a train set duration of about 3 minutes by speaker (which is quite close to the duration available for main condition of NIST evaluation campaigns) and to maximize the number of speakers (despite its large size, the *ESTER-Train* corpus cannot provide a large number of speakers with higher value of A). Considering the test segments, two sizes are investigated (T parameter): T equal to 200 and T equal to 500. As previously, this test segment sizes are chosen according to the ESTER corpus availability in terms of speakers (especially regarding the *ESTER-Test* corpus, which represents 10 hours of speech only). If some speakers present in the *ESTER-Test* corpus have a limited amount of data (less than 200 or 500 words or just above to provide a unique test segment), others may provide several test segments. Regarding the identification task, a first set of evaluations is carried out with all the T-size segments available per speaker (*T200-X* and *T500-X*), whereas a second set of evaluations relies on the first T-size segment per speaker only (*T200-1* and *T500-1*). Four subsets of test data are consequently available for the experiments and are summarized in Table 1.

Protocols		T200-X	T200-1	T500-X	T500-1
# Models (A1000)		19	19	10	10
Ident.	# Tests	103	19	36	10
Verif.	# Client Tests	103	/	36	/
	# Impost. Tests	8233	/	1313	/

Table 1. Description of the experimental protocols used for both the identification (*ident.*) and verification (*verif.*) tasks in terms of number of speaker models, number of tests (*ident.*) and number of client and impostor tests (*verif.*)

3.5. Results

Identification task

Table 2 presents different results obtained for the identification task.

Protocols		T200-X	T200-1	T500-X	T500-1
Correct Ident.	Lexical	7.77%	15.79%	11.11%	20.00%
	Syntactical	12.62%	5.26%	27.78%	40.00%

Table 2. Correct identification rates for various protocols, depending on the modeling type: lexical or syntactical. (T200-X: 103 tests, T200-1: 19 tests, T500-X: 36 tests, T500-1: 10 tests)

At first sight on table 2, the syntactical modeling appears as rendering better scores than the lexical one, excepted for the T200-1 protocol. But, if it seems that syntactical modeling and first-segment tests return better results than lexical modeling and all-segments tests, the T200-1 protocol with syntactical modeling makes us wondering if these two factors are really determinant. In order to answer this question, another criterion can be observed and may inform us the information quantity conveyed: ranks of target models compared to impostor models. Thus, for each test, we calculate the rank of the target model, from 1, if the target speaker is correctly identified, to 19 for T200 evaluations and 10 for T500. Table 3 shows the average rank of the target models for the 4 sets of tests. Curves 1 and 2 represent rank distribution: probability for the target model to be in n first ranks. Looking at the rank criterion on table 3, and figures 1 and

Protocols		T200-X	T200-1	T500-X	T500-1
Average Rank	Lexical	9.69	10.21	5.67	4.60
	Syntactical	6.75	7.05	3.67	3.00

Table 3. Average ranks of target models for various protocols, depending on the modeling type: lexical or syntactical

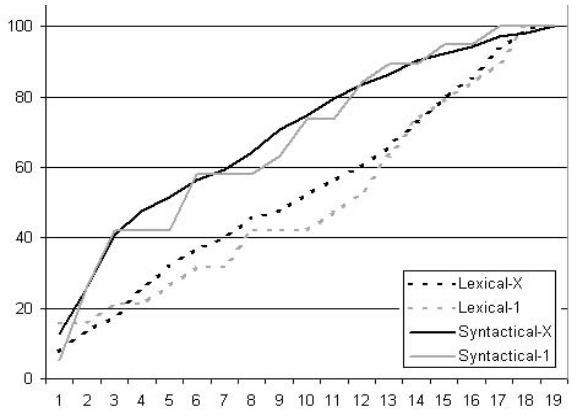

Fig. 1. Cumulative rank distribution of speakers depending on the modeling type: lexical or syntactical, T200-1 and T200-X protocols

2, we can see that first-segment tests obtain better results than all-segments tests for 500 words, but also worth results for 200 words. Finally the most important to remark in the closed-set identification tasks on broadcast news is that, whatever the protocols, syntactical modeling returns better results than lexical modeling. We can even note the lexical modeling leads to the similar scores that random tests would get.

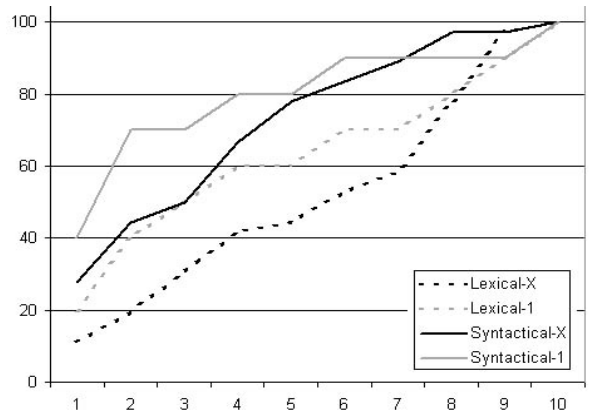

Fig. 2. Cumulative rank distribution of speakers depending on the modeling type: lexical or syntactical, T500-1 and T500-X protocols

Verification task

Figure 3 gives the DET performance curves of both the lexical and syntactical approaches for the speaker verification task. The DET curves are proposed for both the protocols T200-X and T500-X. From this figure, it can be pointed out that syntactical approach outperforms the lexical one: averaged 35% EER for the syntactical approach against 50% for the lexical one. Moreover, similar behaviors may be observed on both the protocols. Despite the small size of the test protocols, in terms of client trials compared with impostor ones, the results reached with the syntactical approach are very interesting and tend to show that syntactical information may carry relevant features for speaker characterization.

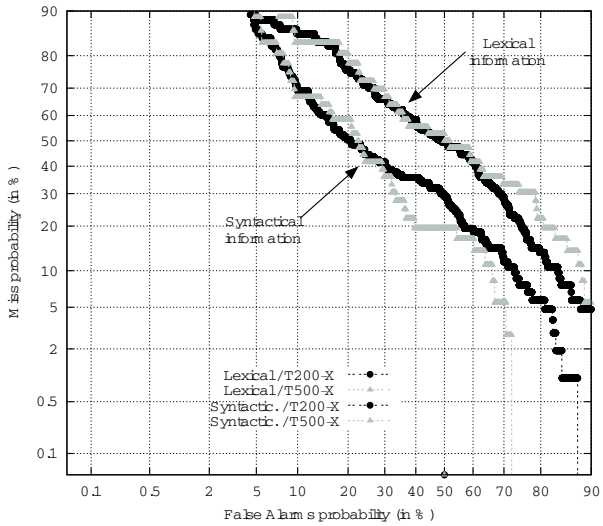

Fig. 3. DET curves for lexical and syntactical systems, T200-X and T-500X protocols

4. CONCLUSION

In this paper, we explored a speaker characterization at two different linguistic levels, the lexical content and the syntactical form, in the context of broadcast news speaker tracking task. We observed that lexical information modeled by n-grams is not useful, in contrary with results obtained on NIST SRE extended task [7]. This result comes partly from the small amount of data available in our case (1000 words for training, 500 or 200 words for testing). The nature of the data, Broadcast News, explains also this difference: it is reasonable to think that lexical information is more related to the topic in our case, when it could be more speaker specific with spontaneous speech data. The syntactical information seems more promising as it allows significant speaker identification (until 40% of correct identification rate) and verification performance (until 35% of EER).
The experiments presented in this paper are relatively preliminary, regarding the small size of the test set (10 hours of BN data) and the use of reference transcriptions. Future works will focus on these points, by using larger database (but it is quite difficult to find large database with a large set of speakers speaking a long time and a correct transcription, particularly for the speaker identity) but also by applying our method on the output of our fully automatic BN rich transcription system. We will also try to apply our syntactical speaker recognition approach for NIST SRE campaign. The main interest is to measure the effect of the nature of the speech (from BN to conversational) on the syntactical information in terms of speaker specificity. A comparison in this context of lexical and syntactical information could be also very interesting.

5. REFERENCES

[1] NIST, Benchmark tests: Rich transcription (RT)., http://www.nist.gov/speech/tests/rt/.

[2] NIST, Benchmark tests: Text retrieval conference (TREC)., http://trec.nist.gov/.

[3] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, G. Gravier, The ester phase II evaluation campaign for the rich transcription of french broadcast news, in: Proceedings of Eurospeech, Lisboa, Portugal, 2005.

[4] D. Istrate, N. Scheffer, C. Fredouille, J.-F. Bonastre, Broadcast news speaker tracking for ester 2005 campaign, in: Proceedings of Eurospeech, Lisboa, Portugal, 2005.

[5] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, Phonetic, idiolectal, and acoustic speaker recognition., in: Odyssey Conference, Chania, Greece, 2001.

[6] N. Scheffer, J.-F. Bonastre, Speaker detection using acoustic event sequences, in: Proceedings of Eurospeech, Lisboa, Portugal, 2005.

[7] G. Doddington, Speaker recognition based on idiolectal differences between speakers, in: Proceedings of Eurospeech, Aalborg, Denmark, 2001, pp. 2521–2524.

[8] J. P. Campbell, D. A. Reynolds, R. B. Dunn, Fusing high- and low-level features for speaker recognition., in: Proceedings of Eurospeech, Geneva, Switzerland, 2003, pp. 2665–2668.

[9] F. Bechet, A. Nasr, F. Genet, Tagging unknown proper names using decision trees, in: 38th Meeting of the Association for Computational Linguistics, ACL'2000, Hong-Kong, 2000.

[10] A. Stolcke, Srilm an extensible language modeling toolkit, in: Proceedings of ICSLP, Denver, USA, 2002.