# Extracting true speaker identities from transcriptions

Yannick Estève, Sylvain Meignier, Paul Deléglise, Julie Mauclair

# Extracting true speaker identities from transcriptions

*Yannick Estève, Sylvain Meignier, Paul Deléglise, and Julie Mauclair*

LIUM - Université du Maine, Le Mans, France

firstname.lastname@lium.univ-lemans.fr

## Abstract

Automatic speaker diarization generally produces a generic label such a *spkr1* rather than the true identity of the speaker. Recently, two approaches based on lexical rules were proposed to extract the true identity of the speaker from the transcriptions of the audio recording without any *a priori* acoustic information: one uses *n-gram*, the other one uses semantic classification trees (SCT). The latter was proposed by the authors of this paper. In this paper, the two methods are compared in experiments carried out on French broadcast news records from the ESTER 2005 evaluation campaign. Experiments are processed on manual and automatic transcriptions. On manual transcriptions, the *n-gram*-based approach can be more precise, but the automatic transcriptions, the SCT-based approach gives significantly the best results in terms of recall and precision.

**Index Terms**: speaker diarization, speaker name extraction

## 1. Introduction

Very large collections of speech data are now available and have to be indexed to allow later retrieval of recorded information. The cost of manual transcription of audio recordings is high, especially when specific indexing is wanted such as the main topic, keywords or the name of the speakers. Automatic rich transcription can be done at a reasonable cost, but the error rate of the systems has to be as low as possible to allow efficient exploitation of their outputs.

The first step to automatically get rich transcription consists in finding the beginning and the end of each homogeneous audio segment which contains the voice of only one speaker; the resulting segments are then clustered by speaker. This step is called diarization in the NIST terminology. Diarization is performed without any prior information: neither the number of speakers, the identities of the speakers, nor samples of their voice are needed. In the literature, the main recent methods are only based on acoustic features [1, 2]. This information allow to increase the accuracy of the next step that is the automatic transcription of the pronounced words.

However, speaker diarization only tags segments with anonymous, automatically-generated identity labels, which are far less useful for multimedia audio indexing than the real identity of the speakers. These labels only allow to determine what are the speech segments pronounced by each person speaking during a show. They are not sufficient to associate a sentence with the true identity of his/her author. Thus, it is very useful to develop effective methods making it possible to find the true author of a sentence.

Currently, there are two main kinds of systems which can be used to associate the true identity of a speaker with the corresponding diarization segments.

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

The first one is based on the analysis of acoustic information. These systems generally rely on automatic speaker recognition methods needing additional samples of the voice of speakers. These samples are used to train acoustic models for targeted speakers [3]: this implies that *a priori* acoustic information is available for each targeted speaker, and of course that all these speakers are already known. This constraint implies a high increase of the development cost of such systems and makes them difficult to manage and to deploy.

The second kind of systems is based on linguistic information. Two different approaches can be followed.

The first approach seems like the one based on acoustic information: it consists in analyzing the sequences of words used by each targeted speaker to characterize his/her manner of speaking [4]. But, this approach presents the same constraint as the approach based on acoustic information: it is necessary to get sufficient *a priori* information about each speaker in order to determine features which characterize him/her.

The second approach based on linguistic information consists in extracting speaker identities directly from speech transcriptions [5, 6, 7]. This approach can be used under certain conditions, especially when speakers have to introduce themselves or have to call the name of other speakers: it seems particularly well suited to broadcast news transcriptions. The main interest of this approach is that it does not require *a priori* information: the true name of the speaker and his localization are generally present in the transcription and can be used to identify this speaker with his/her full name.

In this article, we will treat the latter approach, which consists in extracting speaker identities directly from speech transcriptions. A first study of this approach based on rules defined manually was proposed in [5]. Last year, two different automatic methods were proposed in order to use transcriptions of broadcast news recordings to assign a real full name to the generic labels provided by a speaker diarization system [6, 7].

In this article, an overview of these two methods is presented, and some experiments on French broadcast news data are presented, leading to a comparison of the two methods.

## 2. Generic approach based on lexical rules

The objective of the various methods proposed in [5, 6, 7] is to extract true speaker identities from transcription and associate them with anonymous labels generated by a diarization system. These methods have some common points which can be used to describe a generic approach.

First, the input data are the same for every method: human or automatic transcription of speech data in association with the result of a diarization process. The transcriptions are processed in order to assign categories to some words or sequences of words. These categories are used to generalize some events and it has be shown in [6] that they improve performances of such

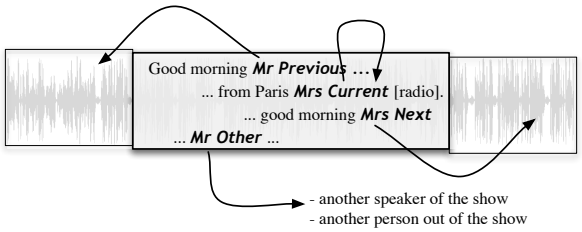


Figure 1: Tags on full names: about whom the speaker is talking?

approach. In [6], this categorization was processed manually. In the experiments presented in this paper, the categorization was automatically processed by the use of an automatic named entity detection system.

When a full name is detected in the transcription a mechanism is activated to determine whether the name refers to the previous speaker, the next one, the current one or another one. Figure 1 illustrates this tagging.

For each detected full name, a local decision has to be made. This leads to a set of local decisions which can associate names with either the speech segment where they were found, or adjacent speech segments.

These local decisions are made by using rules applied on the lexical context of a detected name. In [5], these rules are manually generated. The nature of the lexical rules constitutes the first difference between the two main automatic approaches described below. Each lexical rule provides a score and a tag which is either *'previous'*, *'current'*, *'next'*, or *'other'*.

After applying the lexical rules, each anonymous label issued from diarization can find itself associated with more than one name. This set of names is obtained by linking each name associated with each segment of the cluster corresponding to this anonymous label.

So, it is necessary to choose the most probable name which will be retained as the true identity of the speaker corresponding to this cluster of segments generated by the diarization process.

Then a global decision has to be taken according to local decisions in order to retain this unique name. This global decision uses the scores provided by the local rules to determine the name having the highest score. Here is another difference between the two methods presented below.

## 3. *n-grams* as lexical rules

In [6], speaker identities are extracted from transcriptions of audio recordings by looking for the *n-gram* context around a speaker name detected in the transcriptions. The author uses up to 5-grams with the preceding words, subsequent words and words including the name.

### 3.1. Local decisions

This *n-gram* context is used to predict if this name corresponds to the name of the previous, current or next speaker. For these three situations, *n-gram* rules are built from a training corpus, and are assigned a probability of being correct: each *n-gram* occurring more than 5 times is considered as a predictive rule, and only rules whose probability exceeds a threshold are kept.

### 3.2. Merging local decisions

When a rule with probability $p_1$ proposes the name $n_x$ for a speaker cluster $s_a$, the score $L(s_a = n_x)$ for the hypothesis that $s_a = n_x$ is increased. So, when several rules propose the same hypothesis $n_x$, their probabilities are combined.

For example, if the hypothesis that $s_a = n_x$ is supported by two rules with probabilities $p_1$ and $p_2$, we have:

$$L(s_a = n_x) = p_{1+2} = 1 - (1 - p_1)(1 - p_2)$$

Moreover, a back-off system is proposed to stop duplicative rules: only *n-grams* with the highest order are applied when they contain (N-i)-grams supporting the same prediction.

More details about this method are presented in [6].

## 4. Semantic classification trees as lexical rules

Another method uses semantic classification trees to automatically learn the lexical rules, previously presented by the authors of the current paper in [7] .

### 4.1. Local decisions

When a speaker name is detected, the lexical context of the transcription is analyzed to take a decision about a possible tag of this occurrence, among the *'previous'*, *'current'*, *'next'*, or *'other'* ones. This analysis is made by using a binary decision tree based on the principles of semantic classification trees (SCTs) [8].

SCTs can be very useful to process natural language: for example, they were used for dialog systems [8]. SCTs are based on the use of regular expressions. Pairs composed of a speaker name occurrence and its lexical context are classified according to the comparison between this context and regular expressions. These pairs are classified into the four tags described above. Figure 2 shows an example of SCT.

An improvement of SCTs consists in adding more global questions to questions based on regular expression. For example, improved SCTs are able to use as a classification criterion the position of the speaker name in the speech turn, *i.e.* if the speaker name appears in a short segment, or at the beginning, at the middle or at the end of a long speech segment.

As shown in figure 2, the chosen tag is associated with a score. This score is the probability that a sample reaching the corresponding leaf was associated to this tag in the training corpus used to build this tree. In [7], a local decision about a tag is retained only if this tag is associated with the best score alone. Furthermore, it should be possible to use a threshold in order to validate a local decision according to its local SCT score.

### 4.2. Merging local decisions

The final speaker name associated to a speaker cluster is the name whose occurrences maximize the sum of values given by the SCT, these occurrences referring to segments associated to this speaker cluster. This simple formula permits to take into account the number of occurrences observed for a speaker name candidate, weighted by the SCT scores.

## 5. Experiments

The objective of these experiments is to compare these two approaches which consist in extracting speaker identities directly from speech transcriptions using either *n-gram* as lexical

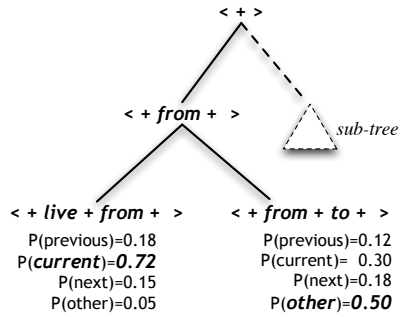


Figure 2: Example of a semantic classification tree

rules [6], or semantic classification trees [7]. These two different automatic methods were proposed in order to use transcriptions of broadcast news recordings: the experiments were made on such data.

### 5.1. Data description

The methods were trained and evaluated with data from the ESTER evaluation campaign. ESTER is the evaluation campaign of French broadcast news transcription systems which started in 2003 and completed in January 2005 [9].

The data were recorded from six radios. They are divided into three corpora: the training corpus corresponds to 81h (150 shows) composed of 8547 segments in which 3297 full names are detected, the development corpus[1] corresponds to 12.5h (26 shows) split into 2294 segments containing 920 full names, and the evaluation corpus contains 10h (18 shows) split into 1417 segments in which 507 full names are detected. The evaluation corpus corresponds to the official ESTER evaluation corpus. This corpus contains two radios which are not present in the training and the development corpora. It was also recorded 15 months after the previous data.

The data were processed using an automatic named entities (NE) detection system. The detected NEs are used to categorize some words or sequences of words: as seen above, in [6] it was shown that using categories to generalize some events improves the performances of the *n-gram*-based approach.

### 5.2. The LIUM NE detection system

The NE detection system used for these experiments was built by the LIUM to participate to an experimental part of the ESTER evaluation campaign on NE detection [9]. It is a rules-based system. Some rules were inferred from the ESTER training corpus, and other ones were developed manually (for example, to produce a grammar to detect dates). Moreover, some lists of first names, last names, cities, countries, *etc.* were injected in the knowledge base of this system.

The NE tagset chosen in the ESTER campaign was made of 8 main categories (persons, locations, organizations, socio-political groups, amounts, time, products and facilities) and over 30 sub categories. For the experiments of this paper, we only used 5 categories: persons, locations, organizations, socio-political groups, and time. Table 1 shows the performance of the LIUM NE detection system on the development corpus of the ESTER evaluation campaign in terms of recall and precision: the NE detection task was only an experimental task in the ESTER campaign; so there was not an official final evaluation.

| EN category | *recall (%)* | *precision (%)* |
|---|---|---|
| person | 98.7 | 92.6 |
| location | 83.9 | 87.9 |
| organization | 86.1 | 84.5 |
| soc.-pol. group | 84.0 | 91.8 |
| time | 96.7 | 97.7 |

Table 1: Performances of the LIUM NE detection system on the manual transcriptions of the ESTER development corpus.

For the experiments presented this paper, all the corpora (training, development, and evaluation) were processed with the LIUM NE detection system. When a NE was detected, the corresponding sequence of words was replaced by the name of the NE category.

### 5.3. Evaluation method

In the framework of speaker identification, the errors consist in identifying a speaker with a wrong identity chosen in a set of known speaker identities. In the presented task, only the public speaker names, those with a full name in the reference, are the clients. The identities of the other ones cannot be found. In our experiment, the public speaker names list contains 1007 names, which are all the names of the known speakers in the training, development and evaluation corpora. The evaluation metrics used to compare the two approaches are the one proposed in [6]: recall and precision. All the proposed results are computed in terms of segment duration as it is done in the NIST evaluations of the speaker diarization, and as it was done in [6].

### 5.4. Training

The *n-gram*-based system was trained with an order *n* equals to 5, and only the *n-grams* observed more than 5 times were kept: these values are the same as the ones used in [6]. We have tested other values, but these ones gave the best results on our development corpus. The SCT-based system was trained by analyzing the lexical context of the detected speaker name: this context was compounded by the 5 words at the left of the name, and by the 5 words at its right. Moreover, the position of the detected speaker name in the speech turn of the current speaker is taken into account to built the SCT, as seen in section 4.1.

### 5.5. Results on manual transcriptions

A first comparison between the *n-gram*-based system and the SCT-based one was made on the evaluation corpus using manual transcriptions. Manual diarization was also used. For both systems, several thresholds (0.1 to 0.9 by step of 0.1) were tested to validate their local decisions, as seen in sections 3.1 and 4.1. Figure 3 shows that the *n-gram*-based system can be the most precise: the precision can reach 100%, but this implies a very low recall of 15.98%. The SCT-based system cannot reach 100% of precision, but for a precision equals to 95.84%, the SCT-based system reaches 42.24% of recall. Notice that the SCT-based system is less sensitive to the variation of the threshold.

[1] it is the official ESTER phase I development corpus merged with the official ESTER phase II development corpus

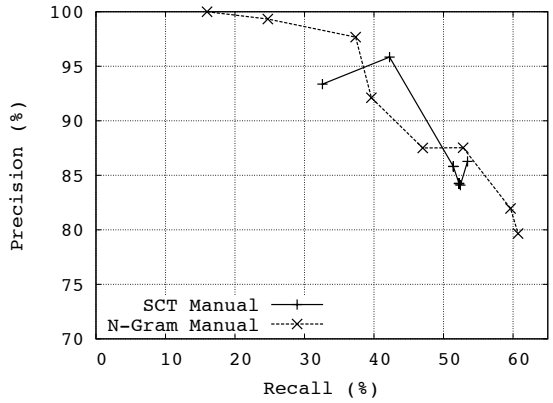


Figure 3: *n-gram vs.* SCT with manual transcriptions.

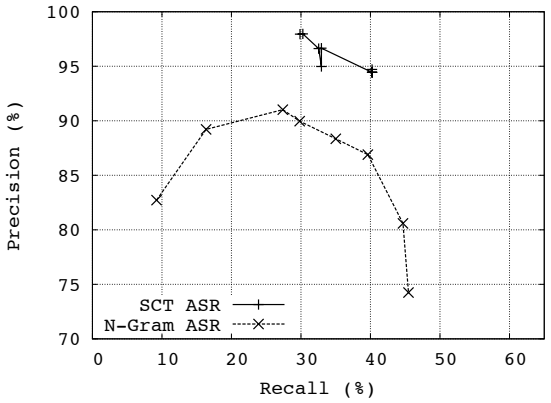


Figure 4: *n-gram vs.* SCT with automatic transcriptions.

### 5.6. Results on automatic transcriptions

Results from using the automatic transcripts instead of the reference information are shown in Figure 4. For this experiments, manual diarization was used. The automatic transcripts were provided by LIMSI [10]: they are the official transcripts submitted by LIMSI during the ESTER evaluation campaign [9]. The LIMSI automatic system recognition offers state-of-the-art performance: these transcripts get a word error rate of 11.9%. Figure 4 shows that the SCT-based system is more robust than the *n-gram*-based system. The SCT-based system outperforms the *n-gram*-based one in terms of precision and recall. While the SCT-based system reaches a precision comprise between 94.44% and 97.97% for a recall comprise between 29.86% and 40.23%, the *n-gram*-based system cannot exceed 91.02% for a recall equals to 27.39%. As with manual transcription, we can observe that the SCT-based system is less sensitive to the variation of the threshold.

## 6. Conclusion

This paper presents a comparative study between two recent approaches to extract true speaker identities from transcriptions. Experimental results shows that the SCT-based method offers better results than the *n-gram*-based one on automatic transcriptions. It is the first time that the SCT-based method is experimented on automatic transcriptions, and it is the first time that the SCT and *n-gram*-based methods are experimented with the use of an automatic named entity detection system.

Future work will focus to integrate confidence measure of ASR outputs into the decision process. This work will tend to develop en entire automatic system, including automatic diarization. Moreover, it should be natural to combine such lexical approaches with more classical acoustic approaches.

## 7. Acknowledgements

The authors would like to thank the SLP Group of LIMSI to have made possible the use of their ASR outputs.

## 8. References

[1] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between GMMs," in *Proc. of ICSLP, International Conference on Spoken Language Processing*, Jeju, Korea, Oct. 2004.

[2] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, vol. 4, pp. 430–451, 2004.

[4] W. Antoni, C. Fredouille, and J.-F. Bonastre, "On the use of linguistics information for broadcast news," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[5] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *Proc. of ASRU, Automatic Speech Recognition and Understanding*, San Juan, Porto Rico, USA, Nov. 2005.

[6] S. E. Tranter, "Who really spoke when? Finding speaker turns and identities in broadcast news audio," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 1013–1016.

[7] J. Mauclair, S. Meignier, and Y. Estève, "Speaker diarization: about whom the speaker is talking?" in *IEEE Odyssey 2006*, San Juan, Puerto Rico, USA, 2006.

[8] R. Kuhn and R. De Mori, "The application of semantic classification trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 449–460, 1995.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[10] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing french broadcast news?" in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.