SPEAKER DIARIZATION: ABOUT WHOM THE SPEAKER IS TALKING ?

J. Mauclair, S. Meignier, Y. Estève

LIUM, Université du Maine
Le Mans, France
{julie.mauclair,sylvain.meignier,yannick.esteve}@lium.univ-lemans.fr

ABSTRACT

The automatic speaker diarization consists in splitting the signal into homogeneous segments and clustering them by speakers. However the speaker segments are specified with anonymous labels. This paper suggests a solution to identify those speakers by extracting their full names pronounced in French broadcast news. A semantic classification tree is automatically built on a training corpus and associate the full names detected in the transcription of a segment to this segment or to one of its neighbors. Then, a merging method permits to associate a full name to a speaker cluster instead of an anonymous label provided by the diarization.

The experiments are carried out over French broadcast news records from the ESTER 2005 evaluation campaign. About 70% show duration is correctly processed for both development and evaluation corpora. On the evaluation corpus, 18.2% show duration is wrongly named and no decision is taken for 11.9% show duration.

1 Introduction

Large collections of speech data are now available but unfortunately, for most of them, without rich transcription. Manual rich transcriptions of audio recordings are high-cost, especially for indexing applications based on specific information like the main topic, keywords, the name of the speaker... Only automatic methods produces rich transcriptions with a reasonable cost, but the error rate due to the performances of the systems must be sufficiently low to be exploited. In this article, the indexing key is the speaker identity.

The first step to automatically get rich transcriptions 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; it is also known as speaker segmentation. The diarization is performed without any prior information: neither the number of speakers, nor the identities of speakers nor samples of their voice are needed. In the literature, the main recent methods are only based on acoustic features [1–4]. The next step consists in transcribing automatically the resulting segments in order to get the pronounced words. Other information can be added as the channel type, the gender of the speaker or the nature of the background.

However, speaker diarization only attributes anonymous labels to segments, whereas the speaker identity is an important criterion for multimedia audio indexing. Speaker identification should be done after the diarization and transcription processes. They are two methods that associate the true identity (full name) of a speaker to the diarization segments:

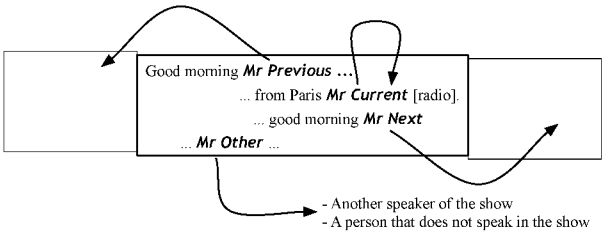

Fig. 1. Tags on full names: about whom the speaker is talking ?

- Acoustic based systems generally rely on automatic speaker recognition methods needing additional samples of the voice of speakers in order to learn acoustic models [5].
- Linguistic based systems extract speaker identities directly from the speech. Speakers often introduce themselves or the next speaker, greet the next or the previous speaker, sign off at the end of their report... The true name of the speaker and his localization are generally present in the pronounced words and can be used to identify speakers with their full name. Compared to the previous method, no speaker voice sample is needed but transcription is necessary.

Recent work carried out on English broadcast news [6, 7], show that a speaker full name occurring in a linguistic context can be used to identify the speaker of the segment with his true name. The linguistic patterns are manually defined in order to tag one of the current, next or previous segment associated to the detected speaker name: "such situations mainly correspond to announcements of who is speaking, who will speak or who just spoke" (*sic*) [7]. They show that the error rate of their tagging process based on manual rules is about 13% and 18% respectively for manual transcriptions and for automatic transcriptions.

We have designed an automatic speaker naming system based on the use of a semantic classification tree which automatically learns such patterns. However, those patterns only provide a local decision for the current segment and the contiguous segments. Then, the system spreads the speaker identity on the entire show. The conflicts are taken into account thanks to the scores provided by the semantic classification tree.

This preliminary study presented in this paper is made to evaluate the relevance of the proposed method. Consequently, only manual diarization and manual transcription references are used here as an input of the system, as it is known that errors coming from automatic diarization and transcription processes reduce the perfor-

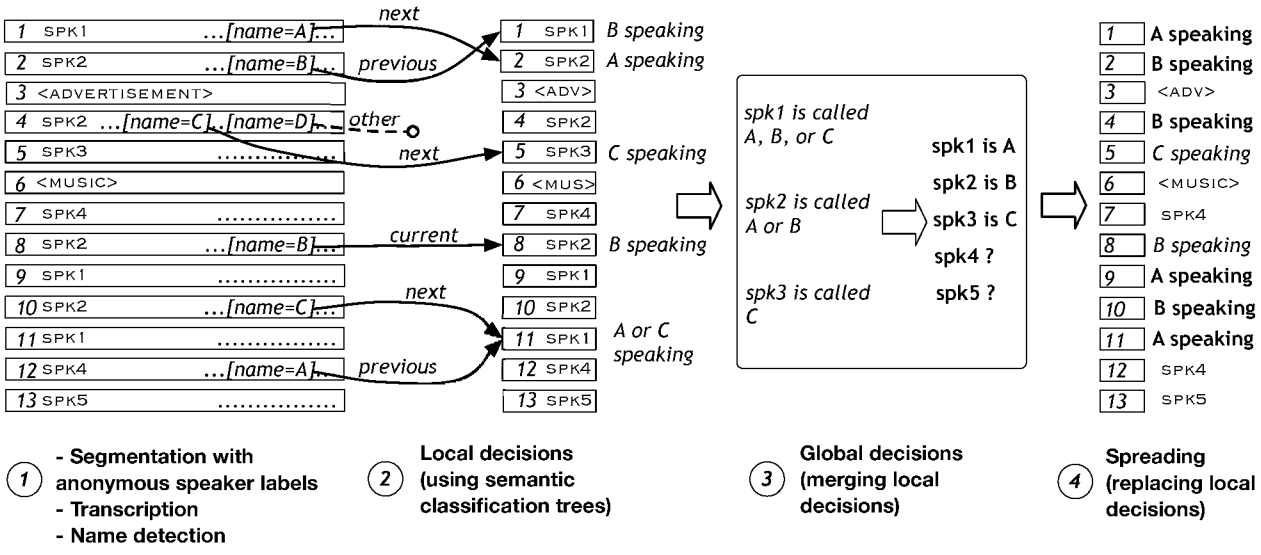

Fig. 2. Speaker identification process

mances of speaker identification based upon a lexical stream (see results of [7]).

Data used for training, development and evaluation are composed by French broadcast news coming from the French 2005 ESTER evaluation campaign [8, 9]. However, the proposed method can easily be applied to English corpora thanks to the full automatic process used for tagging the segments and for speaker naming.

This paper is organized as follows. Section 2 presents the speaker information used in the study. Section 3 describes the method and section 4 the experiments carried out on ESTER corpora.

2 Speaker information

2.1 Client identity

Broadcast news speakers are mainly composed of public persons like journalists, politicians, artists or sportsmen. This population is easily recognizable: their full names are well known, they are present in several broadcast news, and they correspond to the main speakers (in terms of speech duration). These speakers are identified by their full names in the ESTER or LDC transcription conventions and they are the speakers to identify in the proposed task.

A list of speaker identities is extracted from the reference transcriptions. Only the names of well-known persons are kept, others are removed. 1007 different full names were extracted from the corpora used in our experiments.

The speaker name detection process relies on this closed list. We have chosen to use the full name instead of the last name to avoid the false detections introduced by the speaker name detection method. Moreover, ambiguity introduced by the use of the partial names (only forename or family name) leads to problems which we will not resolve in this paper.

2.2 Tags on full name occurrences

Full name located in a show and its context give information to identify the speaker or its neighbor speakers. In fact, a full name in a segment can be associated to one of the following four tags: *current*, *next*, *previous* and *other*. Those tags determine if the detected full name refers to the speaker of the *previous* speech segment, of the *current* one, of the *next* one, or if this full name does not refer to such speakers (see figure 1). In fact, the *other* tag corresponds to the default tag when the full name cannot be attributed to one of the three first tags.

3 Method

Given a set of segments and their transcriptions, we suggests two main processing steps to associate a full name to an anonymous speaker label provided by the diarization process (figure 2, part ①):

1. **Lexical context analysis into each speech segment containing a full name** (figure 2, part ②): this first step processes each full name detected in the transcription of a speech segment. It determines if this full name refers to the previous, current, next or another speaker. Only the segments very close to a full name detected in the transcription can be associated to this full name. Moreover, some segments can be associated to different full names: processes on detected full names are made without cooperation and can provide antagonistic results for the same segment.
2. **Speaker naming** (figure 2, part ③): the second step consists in merging previous hypotheses to assign a full name to an anonymous speaker label. This step spread this assignment to all the segments tagged with this same anonymous speaker label: new results replace first hypotheses obtained at segment level from the previous step (figure 2, part ④).

3.1 Lexical context analysis

When a full name is detected, the lexical context of the transcription is analyzed to take a decision about a possible tag of this full name. This tag helps for naming speaker of contiguous speech segments. This analysis is made by using a binary decision tree based on the principles of semantic classification trees (SCTs) [10].

Semantic Classification Tree

SCTs can be very useful to process natural language. For example, they were used for dialog systems [10], for hierarchical *n-gram* language models estimation [11], or for unknown proper names tagging [12]. SCTs are based on the use of regular expressions. Pairs composed of a full name occurrence and its lexical context are classified according to the comparison between this context and regular expressions. Our aim is to classify these pairs into four tags: *previous*, *current*, *next* and *other* (see leaves in figure 3).

SCT training

During the SCT building process, each node is associated to a regular expression containing words and special characters (<, > and +). < (resp. >) refers to the begin (resp. the end) of a sentence while + refers to any sequence of words. For example, the regular expression < + *from* + > matches every sentence containing the word *from*, while < + *live* + *from* + > matches every sentence containing the words *live* and *from* appearing in this order. Figure 3 shows a very little part of such classification tree.

The SCT building process has to choose for each node the regular expression which minimizes an impurity criterion. For each level in the tree, this building process can only add one word to the current regular expression. The impurity criterion permits to evaluate the degree of determinism associated to a node: lower this impurity criterion is, more the classification should be reliable.

At the end, each leaf is able to give a probability to each possible tag (here: *previous*, *current*, *next* and *other*) for a full name according to the lexical context of the segment where it was detected.

Local decisions

For a given full name occurrence o detected into a lexical context $W_s(o)$ associated to the speech segment s, SCT is able to give the probability $P(t|W_s(o))$ for each possible tag t from tag set $T = \{previous, current, next, other\}$.

Let us define the tag $\delta(o) \in T$ associated to the full name occurrence o in the speech segment s. This tag is given by the formula:

$$\delta(o) = \underset{t}{\operatorname{argmax}}\, P(t|W_s(o)) \tag{1}$$

In our actual approach, beyond the four possible tags for $W_s(o)$, only tag $\delta(o)$ is taken into account for the process continuation. Furthermore, if more than one tag have a probability value equals to $\max_t P(t|W_s(o))$, no local decision is retained.

Let us define the value $\Gamma(o)$ as:

$$\Gamma(o) = P(\delta(o)|W_s(o)) \tag{2}$$

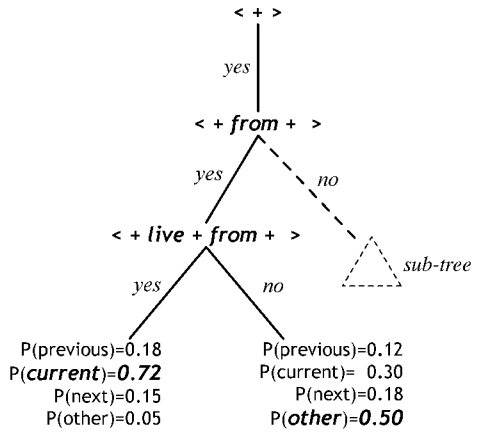

Fig. 3. Example of branch and leaves of a semantic classification tree: for each leaf, a probability value is associated to each tag.

3.2 Speaker naming

The goal of this work is to bind a full name with an anonymous speaker label when it is possible. We note ψ an anonymous speaker: we want to find the real full name $N(\psi)$ of this speaker.

Each segment of speech is associated to its speaker represented by an anonymous speaker label (for example in figure 2, segment 1 is associated to SPK1, as well as segments 9 and 11; segment 2, 4, 8, 10 are associated to SPK2, ...). Moreover, using a semantic classification tree on full names detected in transcriptions of speech segments, a list of full names corresponding to possible speakers for some segments is available (figure 2, part ②).

Merging SCT decisions

Let be K the set of all the full names of the client speakers.

Let be ν_ψ the set of the different full names associated by local SCT decisions to at least one segment pronounced by ψ: ν_ψ is the list of full name candidates for ψ and $\nu_\psi \subset K$.

Let us define the function $\nu(o)$ which associates an occurrence o of the full name n to this full name n. In this case, we have: $\nu(o) = n$.

At last, let us define the set Ω_ψ of occurrences o which refer by local SCT decisions to segments pronounced by ψ.

We propose to find the full name $N(\psi)$ of the speaker ψ using the following formula:

$$N(\psi) = \underset{n \in K}{\operatorname{argmax}} \frac{\sum_{(\nu(o)=n) \wedge (o \in \Omega_\psi)} \Gamma(o)}{\sum_{o \in \Omega_\psi} \Gamma(o)} \tag{3}$$

$$= \underset{n \in K}{\operatorname{argmax}} \sum_{(\nu(o)=n) \wedge (o \in \Omega_\psi)} \Gamma(o) \tag{4}$$

So, the full name associated to a speaker label is the full name whose occurrences maximize the sum of values given by the SCT about these occurrences referring to segments associated to this

	Train	*Dev*	*Eva*
# Shows	150	26	18
# Channels	5	5	6
duration (h)	86	12.5	10
# Segments	8547	2294	1417

Table 1. Corpus information: Train, Development & Evaluation from French broadcast news ESTER evaluation campaign.

speaker label. Notice that as explained in section 3.1, only values associated to valid local decisions are kept. This simple formula permits to take into account the number of occurrences observed for a full name candidate, weighted by the SCT scores.

4 Experiments and results

4.1 Data

Corpora

The methods are trained and evaluated with data from the ESTER evaluation campaign. ESTER is an evaluation campaign of French broadcast news transcription systems which started in 2003 and completed in January 2005 [8, 9]. This evaluation campaign was organized within the framework of the TECHNOLANGUE project funded by the french government under the scientific supervision of the AFCP[1] with the DGA[2] and ELDA.

The data were recorded from six radios: *France Inter, France Info, RFI, RTM, France Culture* and *Radio Classique*. The data are divided into three sets; only the two first ones are annotated[3]. Shows (10 minutes up to 60 minutes) from those two first sets contain few silence, music and advertisements comparing to the LDC English broadcast news corpus [13]. The majority of the shows contains prepared speech like news and few conversational speech like interviews. Only 15% of the corpus is narrow band speech. Those data are split in three corpora (described on table 1):

- The training corpus called *Train* corresponds to 81h (150 shows) composed of 8547 segments in which 3297 full names are detected.
- A development corpus[4], denoted *Dev*, corresponds to 12.5h (26 shows) split into 2294 segments containing 920 full names.
- An evaluation corpus, denoted *Eva*, contains 10h (18 shows) split into 1417 segments in which 507 full names are detected. *Eva* corresponds to the official ESTER evaluation corpus. This corpus contains two radios which are not present in the training corpus. It was also recorded 15 months after the previous data.

Table 2 shows the *a priori* probabilities of the four full names tags computed on the reference manual transcriptions. In both cases, the *next* tag is the most frequent one (between 45% and 49%) and the *current* tag is the least frequent one (between 5% and 7%).

[1] AFCP: Association Francophone de la Communication Parlée

[2] DGA: Délégation Générale de l'Armement

[3] they are officially denoted Phase I and Phase II

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

	Train	*Dev*	*Eva*
# detected Full name (*)	3297	920	507
Previous	14.3%	12.6%	18.6%
Current	7.2%	7.1%	5.3%
Next	46.0%	45.3%	49.3%
Other	32.5%	35.0%	26.8%

Table 2. Statistics of full name tags on training, development & evaluation corpora computed over the manual reference.
- (*): the number of speaker full name detected in the corpus.

Preparing the corpora

Transcriptions provided by the corpora are designed for diarization or transcription tasks. References (rich transcriptions) have to be transformed and adapted to be used with a semantic classification tree and to evaluate experiment results. These adaptations are:

- The definition of the four full name tags supposes that the previous and the next speakers are different from the current one. The segmentation must rely on speaker turns and does not rely on sentences (mostly separated by breath and silence) as it was done in the manual transcription. So, the contiguous segments from the same speaker are merged to obtain a segmentation based on speaker turns.
- The information about the four tags are needed during training and scoring phases. We tagged the reference corpus automatically by extracting speaker full names in the speech. Each full name is compared to the speaker full names attached to the segment and its contiguous neighbors. Thus, this automatic task is not checked manually and we suppose that all speaker identifications are correct.
- In the reference transcription, sentences contain more information than those produced by an automatic system. Transcriptions are then normalized to be as close as possible to the ones made by an automatic transcription system. For example, all the punctuations are removed, the upper case are removed, and so on.
- In the same manner, the definite articles (*le, la, les*) and the indefinite articles (*un, une, des*) are removed from the sentences. We believe that they are not informative.
- To generalize the training examples during the building of the tree, each speaker full name is replaced by a generic label.
- The semantic classification tree learns the regular expressions according to the words in the left and right contexts of a speaker full name occurrence. No more than only 40 words around the speaker full name are kept: at most 20 words on the left and at most 20 words on the right. The number of words on the left and on the right was fixed over the *Dev* corpus in order to maximize the number of true local detection of the four tags.

SCT training parameters

The semantic classification tree is tuned on the development corpus. The main parameters for the training are the Gini criterion [14] as the impurity criterion and the size of the leaves. The expansion of the branches stops when the Gini criterion is not reduced or when the current node is associated to less than five sequences of words.

4.2 Segment speaker tagging

	Train	*Dev*	*Eva*
# detected full name	3297	920	507
Tagged	94.51%	94.78%	97.23%
Correctly tagged	88.25%	76.49%	68.76%
Previous	88.98%	71.67%	82.98%
Current	94.76%	90.14%	85.71%
Next	89.32%	80.67%	75.29%
Other	84.87%	68.94%	50.32%

Table 3. Scores of local decisions using the semantic classification tree on training, development & evaluation corpora.
- Tagged: rate of detected full names for which a full name tag is proposed using the local decision rule.
- Correctly tagged: rate of detected full names that are correctly tagged.
- Previous (resp. for the other tags): rate of detected full names that are correctly tagged by previous tag.

The semantic classification tree which provides the results on table 3 was built with the training corpus. The table shows the results of the local decisions taken over each segment containing a detected full name on the *Train*, *Dev* and *Eva* corpora. The first column shows the scores of the train data used as a test corpus. The second and third column report the results on *Dev* and *Eva*.

94% detected full names on *Dev* and 97% on *Eva* are tagged by one of the four full name tags. The correct tagging rate is above 76.4% on *Dev* and only 68.7% on *Eva*: these values can be considered as the precision of the local decision method on each corpus.

The lowest result for *Eva* ($\sim$8% less) can be explained by the presence of two new stations and which are not present in the training and development corpora. The *Eva* data were also recorded 15 months later. About 6% detected speaker full names are untagged as well as in the training corpus.

The results for the *other* tag are the weakest. This tags seems to be associated to more various lexical contexts than the others. In this case, the names can be associated to distant (not contiguous) segments or even to people not intervening in the show. Nevertheless, the impact of this results is low as this tag is not taken directly into account in the naming process.

By always simply choosing the tag having the strongest prior probability (see table 2), we will only reach a score of $\sim$45.3% on *Dev* corpus (respectively $\sim$49.3% for *Eva*). With the method proposed above, $\sim$76% correct tagging rate for *Dev* is observed ($\sim$68% for *Eva*). These results show that the semantic classification tree is well adapted to this task, permitting to exploit them in the speaker naming process, as shown below.

4.3 Speaker naming

Local decisions on the segments are merged to associate one full name to all the segments pronounced by the same speaker (see section 3.2). The detailed results of this second step are reported in table 4.

Evaluation method

The input of the system is based upon the manual transcription references: the diarization (anonymous speaker labels), segmentation in speech/non speech and transcriptions errors cannot exist. The reference and hypotheses segment boundaries are equal, only the speaker names differ.

Speakers	Naming	*Train*	*Dev*	*Eva*
Client	Correct	63.68%	64.82%	66.35%
Client	Wrong	3.19%	5.48%	14.36%
Client	Unnamed	15.68%	18.19%	11.91%
Not client	Correct (unnamed)	15.50%	7.54%	3.59%
Not client	Wrong	1.95%	3.98%	3.79%
	Total	100%%	100%	100%

Table 4. Speaker naming: detailed results on training, development & evaluation corpora (all the rates are computed in terms of duration).
- Speakers: This defines the two categories of speakers in the reference, those which are the clients of the application (public speakers with a full name) and the others.
- Naming: corresponds to the correct and wrong naming. Unnamed is the case where the process is not able to propose a full name.

In the framework of speaker identification, the errors consist in identifying the 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 others cannot be found.

There are errors when the process gives a non-client speaker a full name and when the process does not give a client speaker (a public speaker) a full name (Table 4 lines 2 & 5).

Moreover, the process cannot propose a name to a client speaker in some circumstances:

- no local decision affects a segment of this client speaker. Either no local decision is taken for the detected occurrences of the full name of this client speaker, or all the existing local decisions are wrong;
- the full name of this speaker is not detected in the transcriptions.

For client speakers, when the hypothesis full name and the reference full name are the same, this is considered as a correct naming (Table 4 line 1). For non-client speakers, it seems reasonable to consider correct not to assign a full name of a client speaker to speech pronounced by a non-public person.(Table 4 line 4).

All the proposed results are computed in terms of segment duration as it is done in the NIST evaluations of the speaker diarization [15].

Comments

The speaker naming process gives a correct decision up to 72% speech duration ($64.82\% + 7.54\%$) over *Dev* corpus and about 70% ($66.35\% + 3.59\%$) over the *Eva* corpus as shown in table 4.

The difference on correct naming rate between the *Dev* corpus and the *Eva* corpus is about 2%, less than the 8% observed for the local decisions in table 3. Even if there are less local decisions in the *Eva* corpus, those decisions are relevant for finding the true full name of a client speaker.

5 Conclusion

In the framework of rich transcription, we propose a full automatic method to identify the speakers by their full names extracted from the transcription.

The process is firstly based upon the use of a semantic classification tree which permits to qualify the detected occurrences of full names: this first step consists in local decisions binding each of these occurrences to a speech segment. Then, the local results are merged to associate a full name to all the segments of a given speaker.

The experiments are carried out over French broadcast news records from the ESTER 2005 evaluation campaign. About 70% show duration is correctly processed for both development and evaluation corpora. On the evaluation corpus, 18.2% show is wrongly named and no decision is taken for 11.9% show.

The main goal is reached: the results validate the proposed method of speaker naming processed on manual diarization and manual transcription. Further work will focus on the use of automatic diarization and transcription in which errors are present.

6 Acknowledgements

The authors would like to thank Frédéric Béchet from LIA (Computer Science Lab of the University of Avignon, France) for making LIA_SCT tool available as an open source project.

7 References

[1] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding (IEEE, ASRU 2003)*, St. Thomas, U.S. Virgin Islands, November 2003, pp. 411–416.

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

[3] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *DARPA RT04 Fall*, Palisades, NY, USA, 2004.

[4] 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*, 2005.

[5] 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.

[6] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, Jeju, Oct 2004.

[7] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," Jeju, Oct 2005.

[8] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference (LREC 2004)*, Lisbon, Portugal, May 2004.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," Lisboa, Sep 2005, pp. 1149–1152.

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

[11] Y. Estève, F. Béchet, A. Nasr, and R. De Mori, "Stochastic finite state automata language model triggered by dialogue states," in *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 2001)*, Aalborg, Denmark, 2001, vol. 1, pp. 725–728.

[12] F. Béchet, A. Nasr, and F. Genet, "Tagging unknown proper names using decision trees," in *38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, October 2000, pp. 77–84.

[13] S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *2004: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2004)*, Toledo, Spain, May 2004.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[15] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," `http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v%14.pdf`, August 2004.