Automatic named identification of speakers using belief functions

Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, Yannick Estève

Automatic named identification of speakers using belief functions

Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, and Yannick Estève

Laboratoire d'Informatique de l'Université du Maine
Avenue Laennec, 72085 Le Mans Cedex, France
{simon.petit-renaud,vincent.jousse,
sylvain.meignier,yannick.esteve}@univ-lemans.fr

Abstract. In this paper, we consider the extraction of speaker identity (first name and last name) from audio records of broadcast news. Using an automatic speech recognition system, we present improvements for a method which allows to extract speaker identities from automatic transcripts and to assign them to speaker turns. The detected full names are chosen as potential candidates for these assignments. All this information, which is often contradictory, is described and combined in the Belief Functions formalism, which makes the knowledge representation of the problem coherent. The Belief Function theory has proven to be very suitable and adapted for the management of uncertainties concerning the speaker identity. Experiments are carried out on French broadcast news records from a French evaluation campaign of automatic speech recognition.

Key words: Speaker identification, speaker recognition, information fusion, belief functions.

1 Introduction

In order to allow later retrieval of recorded information, large collections of audio documents have to be indexed. The system presented in this paper focuses on speaker identification by their full name in audio documents. The speaker identity detection is composed of several steps and is in most cases subject to uncertainty and confusion. The first step to automatically get audio documents indexing consists in detecting speakers turns and regrouping those uttered by the same speaker. It is generally based on a first stage of segmentation that consists in partitioning the regions of speech into homogeneous audio segments which contains ideally the voice of only one speaker, followed by a clustering stage that consists in giving the same label to segments uttered by the same speaker. A speaker turn starts when a speaker is starting to speak and ends when another speaker is starting to speak, or a song or advertising is starting. Speaker turns are regrouped by class of same but *anonymous* speakers. The next step is to automatically transcribe the content of speaker turns into words and is complemented by an annotation for some words as "named entities" or categories.

Some words are particularly identified as "PERSON". The more promising way to identify speakers by their *real* full name consists in extracting them from the automatic speech recognition system (ASR) [1, 9, 5, 6]. The general principle is to determine if a detected named entity as a "PERSON" refers to a speaker of the document or to a person who does not speak in the document.

Our article takes place in this framework. The system we developed in [5, 6] uses uttered full names to assign them to anonymous speakers from identified speaker turns. The principle is to assign one of these four labels: *"current turn"*, *"previous turn"*, *"following turn"* or *"another person"* to each detected full name. But this approach ignored the potential conflict information on the speakers within a same speaker turn. In this paper, we propose to improve the consistency of the system and to better combine the various information on the potential speakers. The formalism of Belief Functions seemed to be particularly suited to managing these conflicts and combining this information.

First, we briefly present the automatic transcription system used, before describing the reference system for speaker named identification. Then, we discuss the shortcomings of this model and the improvements of our model using belief functions. Finally, we propose metrics for evaluating such systems, and we comment on the results obtained on the ESTER I evaluation campaign [3].

2 Speaker identification based on a transcription system

2.1 Transcription system

The main hypothesis initially proposed in [1] assumes that a detected full name in a speaker turn allows to identify the current turn or a directly contiguous turn (previous or following turn). However, some full names identify farther speaker turns or persons that are not involved in the document. The used identification method is based on previously transcribed and enriched documents. This transcript needs to cut the document into segments which are then classified in anonymous speakers. These segments, grouped into speaker turns are transcribed and the named entities are annotated. The LIUM transcription system is described in [2]. During the ESTER 1 (phase II) evaluation campaign in 2005 [3], for the transcription task, our system was ranked second. This system achieves 20.5 % of word error rate on the evaluation corpus.

2.2 Semantic classification trees

The speaker identification method uses a binary decision tree based on the principle of semantic classification trees (SCT) [4]. A SCT automatically learns lexical rules from full names detected in the training corpus, with the left and right surrounding words. A SCT is used for each occurrence of full names detected in the transcripts. This tree allows to associate to each occurrence of a full name the probability to correspond to one of the four envisaged hypotheses: *"current turn"*, *"previous turn"*, *"following turn"* or *"another person"*. These probabilities are determined during the learning of the tree and reflect the observed cases in the training corpus.

2.3 Reference combination method

The final goal of the system is to assign a full name to each anonymous speaker. Let us recall here the combination method of information provided by the trees that have been proposed in [5]. Let $\mathcal{E} = \{e_1, \ldots, e_I\}$ denotes the *closed* set of full names hypotheses to assign to a speaker. These candidates come from an exhaustive list of possible speakers known by the system. The set $\mathcal{O} = \{o_1, \ldots, o_J\}$ corresponds to the successive occurrences of full names detected in the transcripts, $\mathcal{T} = \{t_1, \ldots, t_K\}$ is the set of the speaker turns in chronological order, and $\mathcal{C} = \{c_1, \ldots, c_L\}$ is the set of anonymous speakers to be labeled. Thus, the main goal is to assign a full name of $\mathcal{E}$ to a speaker of $\mathcal{C}$. Each speaker c_l may involve one or several times in a broadcast, that corresponds to several speaker turns: $c_l = \{t \in \mathcal{T} \mid c_l \text{ is the speaker of turn } t\}$. In a same turn, several occurrences of full names may be detected. For each occurrence of a full name o_j (for $j = 1, \ldots, J$) detected in a given speaker turn t_k, let us define by $P(o_j, t_k)$ the probability that o_j is current speaker. Thus, $P(o_j, t_{k-1})$ and $P(o_j, t_{k+1})$ represents the probability that o_j is respectively the speaker of the previous and the following turn. By hypothesis, the probability that o_j is another speaker is: $1 - \sum_{r=-1}^{1} P(o_j, t_{k+r})$. At this stage, a filter is made by the comparison of genders: if the gender of the full name e_i and the speaker c_l are different, the corresponding occurrence is ignored. Let $g(e_i)$ and $g(c_l)$ be the respective gender (female, male or unknown) of e_i and c_l. The speaker gender is detected by the acoustic segmentation and classification system with high reliability and the full name gender is determined by the first name (generally without ambiguity) from a linguistic base of first names.

In [5], to assign a full name e_i to a speaker c_l, we have computed a "score" for each full name e_i, denoted as $s_l(e_i)$. This score is no more a probability, but is simply a sum of probabilities concerning the speaker turns of c_l and taking gender constraints into account:

$$s_l(e_i) = \sum_{\{(o_j,t) | o_j = e_i,\ t \in c_l, g(e_i) = g(c_l)\}} P(o_j, t) \tag{1}$$

2.4 Decision process

The goal is now to assign a full name e_i to each speaker c_l. Let $f : \mathcal{C} \rightarrow \mathcal{E}$ be the assignment function of full names to speakers. The principle of our solution, proposed in [5], is actually to find a coherent matching between full names and speakers. Let $\mathcal{D} = \{c_l \in \mathcal{C} \mid \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$, be the set of speakers with no potential candidate. Several strategies may be used to sort the competing speakers c_l for a given full name e_i. The more natural way seems to choose the full name which has the maximum score for a given speaker e_i (if there is at least one non-null score). Let us define the rule $\mathbf{R_1}$ by:

$$\begin{gathered} \forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{s}_l(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonymous}. \end{gathered} \tag{2}$$

An issue is that the same full name e_i^* may be assigned to several speakers. We proposed to reorganize the sharing of full names among speakers. Let the coefficient β_{il} define the relative score of e_i among all the possible candidates for assignment to c_l: $\beta_{il} = \frac{s_l(e_i)}{\sum_{q=1}^{I} s_l(e_q)}$ if $c_l \notin \mathcal{D}$ and $\beta_{il} = 0$ if $c_l \in \mathcal{D}$. A concrete example is given in Table 1. The full name "Jacques Derrida" has been assigned to three different speakers from the decision rule in Equation 2. In this example, c_{13} has the best score, and "Jacques Derrida" should be assigned to c_{13} ; but the score represents only 35% of the total scores among all the possible candidates for c_{13}, whereas the score for c_{15} represents 80% of total scores. Then, for the final decision, we have proposed to use the product of score $s_l(e_i)$, by coefficient β_{il} (rule $\mathbf{R_2}$):

$$\mathrm{SC}_l(e_i) = s_l(e_i)\beta_{il}. \tag{3}$$

Finally, in the same example, "Jacques Derrida" is assigned to c_{15} and the speakers c_{13} and c_{14} will be labeled with other full names. The algorithm is iterative:

Table 1. Example of an initial multiple assignment

Speaker	Full name e_i^*	$s_l(e_i^*)$	β_{il}	$\mathrm{SC}_l(e_i^*)$
c_{13}	Jacques Derrida	**8.58**	35%	3.00
c_{14}	Jacques Derrida	1.67	56%	0.94
c_{15}	Jacques Derrida	4.94	**80%**	**3.95**

all the full names are taken *a priori* into account and sorted according to their score $\mathrm{SC}_l(e_i)$. First, the full name with the maximum score (denoted e_i^*) is chosen, and if several speakers are associated to the same e_i^*, then e_i^* is assigned to the speaker whose score $SC_l(e_i^*)$ is maximum. Then, all chosen full names are deleted from the list of speakers that are not yet assigned in this first iteration. In a second iteration, remaining full names are examined in the same way for the remaining speakers and so on, until all speakers are assigned, or their list is empty. Table 2 shows the result of this algorithm for the preceding example.

Table 2. Example of the decision process with two iterations (decision in bold type, scores in parenthesis).

Speaker	e_i^* (1st iteration)	2nd iteration
c_{13}	J. Derrida (3.00)	**N. Demorand** (0.25)
c_{14}	J. Derrida (0.94)	**A. Adler** (0.56)
c_{15}	**J. Derrida** (3.95)	-
c_{16}	**O. Duhamel** (1.15)	-

2.5 Drawbacks of the combination method

The combination method described above has several serious drawbacks, even though it has yielded good results [5]. First, the concept of score is difficult to interpret, the quantities obtained in Equations 2 and 3 do not represent a degree of confidence, or a probability that a full name is a given speaker. They lead to a lack of clarity of the decision. Equation 3 represents a compromise that is difficult to justify.

But the main drawback concerns the lack of uncertainty management in the combination method: particularly, conflict information in a given speaker turn is not taken into account. The available information is not correctly combined as a whole. No link is made between the different information provided by the classification tree, particularly when a same speaker pronounces several different full names and can therefore lead to erroneous results. Table 3 presents an example of a speaker turn t_k where 8 occurrences are detected. The probabilities correspond to the next speaker turn t_{k+1}, who is a male. A female full name is therefore eliminated and two full names are rejected because they do not belong to the list. Some occurrences are redundant, because they correspond to a repeated full name and only one occurrence has a relative high probability. Two full names are still competing and they represent a significant incompatibility. These full names have high scores: Jean-Claude Pajak (1.25) and Jacques Chirac (0.87). These scores are close to those obtained if we had some information without ambiguity, for example a turn with only one occurrence with a high probability. This example highlights the fact that this method does not take into account the contradictory information provided by some speaker turns. A probabilistic formalism based on conditional probabilities could be considered for this kind of situation, but the lack of *a priori* information makes this type of modeling difficult. Even though the classification tree outputs are probabilistic, belief theory seemed more appropriate and less restrictive, particularly in the flexibility of its use.

Table 3. Score contribution in a speaker turn (t_{k+1} is a male).

Occurrence o_j	gender	belongs to the list	$P(o_j, t_{k+1})$	score
Oscar Temaru	*M*	*No*	*0.29*	–
Hamid Karzaï	*M*	*No*	*0.29*	–
Jacques Chirac	M	Yes	0.29	0.87
Jacques Chirac	M		0.29	
Jacques Chirac	M		0.29	
Jean-Claude Pajak	M	Yes	0.29	
Jean-Claude Pajak	M		**0.96**	1.25
Véronique Rebeyrotte	*F*	*Yes*	*0.29*	–

3 Belief functions for speaker recognition

The contribution of this article lies in the combination process of different information, especially from the classification tree.

3.1 Belief function theory

In this section, we briefly recall some notions of the belief function theory [7, 8]. In this article, we adopt the point of view proposed by Smets: the Transferable Belief Model (TBM) [8]. The aim of this model is to determine the belief concerning different propositions, from some available information. Let Ω be a finite set, called frame of discernment of the experience. The representation of the uncertainty is made by the means of the concept of belief function, defined as a function m from 2^{Ω} to $[0,1]$ such as $\sum_{A\subseteq\Omega} m(A) = 1$. The quantity $m(A)$ represents the belief exactly allowed to proposition A. The subsets A of Ω such as $m(A) > 0$ are called the *focal elements* of m. One of most important operations in the TBM is the procedure for aggregating operator to combine several belief functions defined in a same frame of discernment [8]. In particular, the combination of two belief functions m_1 and m_2 "independently" defined on Ω using the conjunctive binary operator $\cap$, denoted as $m' = m_1 \cap m_2$, is defined as [8]:

$$\forall A \subseteq \Omega,\ m'(A) = \sum_{B\cap C=A} m_1(B)m_2(C). \tag{4}$$

Repeatedly, we may define the combination of n functions $m_1, \ldots, m_n$ on Ω by: $m = m_1 \cap \ldots \cap m_n$. Once a belief function m is defined, it is possible to transform it into a probability distribution particularly for decision aspects. One of these, called *pignistic* probability and denoted by P_m, is defined for all $\omega \in \Omega$ as [8], if $m(\emptyset) \neq 1$:

$$P_m(\{\omega\}) = \sum_{A\subset\Omega} \frac{m(A)}{|A|(1-m(\emptyset))}\delta_A(\omega), \tag{5}$$

where $|A|$ denotes the cardinality of A, $\delta_A(\omega) = 1$ if $\omega \in A$ and $\delta_A(\omega) = 0$ if $\omega \notin A$.

3.2 Definition of belief masses

In this article, we propose to improve the system described in [5] by taking into account the *coherence* of the whole information provided by *contiguous* speaker turns. As we have seen before, in a *same* turn, several occurrences corresponding to different full names may be detected.

First, we focus on a turn t_k with n_k occurrences and owing to speaker c_l. Let n_{k+r} be the number of occurrences for the previous turn ($r = -1$) and the following one ($r = 1$). Let $\{o^k_{j,r}\}$, with $r = -1, 0, 1$ and $j = 1, \ldots, n_{k+r}$, be the occurrences of the detected full names in these three turns. Each occurrence $o^k_{j,r}$, corresponding to a label e_i, represents some knowledge concerning the speaker

of the turn t_k that can be described by a simple support belief function $m^{jr}_{t_k}$ on $\mathcal{E}$, focused on e_i and $\mathcal{E}$:

$$\begin{cases} m^{jr}_{t_k}(\{e_i\}) = \alpha_{ij} P(o^k_{j,r}, t_{k-r}) \text{ si } o^k_{j,r} = e_i \\ m^{jr}_{t_k}(\mathcal{E}) = 1 - \alpha_{ij} P(o^k_{j,r}, t_{k-r}), \end{cases} \tag{6}$$

where $\alpha_{il} \in [0,1]$ is a confidence measure of gender compatibility between e_i and c_l. If the genders are known with certainty, $\alpha_{il} = 0$ if $g(e_i) \neq g(c_l)$ and $\alpha_{il} = 1$ if $g(e_i) = g(c_l)$. If the first names are ambiguous (like Dominique in French) or unspecified, or if the speaker gender is uncertain, $\alpha_{il} \in]0,1[$ is estimated from a database of first names and the training corpus. Table 4 presents the belief function concerning the speaker of turn t_{k+1} in the example seen in section 2.5. The belief mass of "Jean-Claude Pajak" is still high while the one of the other candidate is very low, and the degree of conflict is important since the mass of the empty set is high.

Table 4. Mass distribution of the belief function in a speaker turn.

Focal elements	$m_{t_{k+1}}(\{e_i\})$
Jacques Chirac	0.018
Jean-Claude Pajak	**0.348**
$\emptyset$	**0.624**
$\mathcal{E}$	0.010

3.3 Combination by speaker

The first combination step consists in aggregating the whole information in a given speaker turn. The combination of the $n_{k-1} + n_k + n_{k+1}$ belief functions focused on the t_k and obtained by Equation 6 is made with conjunctive *non normalized* Dempster rule (Equation 4), in order to ensure associativity and commutativity of the combination: we obtain a belief function m_{t_k} that represents a more synthetic knowledge of speaker identity provided in turn t_k, defined by:

$$m_{t_k} = \bigcap_{r=-1}^{1} \bigcap_{j=1}^{n_{k+r}} m^{jr}_{t_k}. \tag{7}$$

The second combination step consists in aggregating the results obtained by each speaker turn for the whole broadcast news. The more relevant and natural consists in keeping on combining all the belief functions focused on the same speaker c_l with the same conjunctive Dempster rule and therefore combining all the belief functions corresponding to the speaker turns t_k of this speaker. Thus, we obtain a global belief function M_l which represents the state of belief concerning speaker c_l for the whole broadcast news, and defined by:

$$M_l = \bigcap_{t_k \in c_l} m_{t_k} \tag{8}$$

3.4 Decision rule

We use a similar procedure presented in section 2.4, but the decision process is simplified and unified thanks to the use of belief functions. We transform the belief function M_l into the pignistic probability P_{M_l} (Equation 5) and we obtain the following rule **R**:

$$\begin{gathered}\forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{P}_{M_l}(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonymous.}\end{gathered} \tag{9}$$

Then, since some full names may initially be assigned to several speakers, we apply the same decision process developed in 2.4, replacing scores SC_l by pignistic probabilities P_{M_l}. If we come back to the proposed example in 2.4, the full name "Jacques Derrida" is again initially assigned to three speakers c_{13}, c_{14} and c_{15} (see Table 5). Finally,"Jacques Derrida" is also assigned to c_{15}, because this speaker has the most important pignistic probability.

Table 5. Decision with two iterations (decision in bold, belief masses in parentheses).

Speaker	e_i^* (1st iteration)	2nd iteration
c_{13}	J. Derrida (0.89)	**N. Demorand** (0.11)
c_{14}	J. Derrida (0.71)	**A. Adler** (0.25)
c_{15}	**J. Derrida** (0.99)	-
c_{16}	**O. Duhamel** (0.88)	-

4 Evaluation of the proposed system

4.1 Data description

The system evaluation are realized on French broadcast news records from the French ESTER 1 phase II evaluation campaign [3]. The data were recorded from 5 French radios and *Radio Télévision Marocaine* and last from 10 to 60 minutes. They are divided in 3 corpora used for the SCT training, the system development and the evaluation: the training corpus of 76h (7416 speaker turns, 11292 detected full names), the development corpus of 30h (2931 speaker turns, 4533 full names) and the test corpus of 10h (1082 speaker turns, 1541 full names). 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.

4.2 Metrics

The results are evaluated comparing the generated hypothesis and the reference. This comparison highlights five cases:

– Identity is correct (C_1): the identity hypotheses corresponds to the correct one in the reference.

- Substitution error (S): the identity hypotheses differs from the one found in the reference.
- Deletion error (D) : no identity is proposed although the speaker is identified in the reference.
- Insertion error (I): an identity is proposed although the speaker is not identified in the reference.
- No identity (C_2): no identity is proposed, and there is no identity for this speaker in the reference.

Among the measures defined in [9, 5], the one that seems to best summarize the results is the global error rate Err computed from these 5 quantities:

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1}. \tag{10}$$

The errors may be computed in terms of duration or in terms of number of speakers.

4.3 System evaluation

During experiments, the system is supposed to know all the full names that may be the speakers. This list is composed 1008 full names (the set $\mathcal{E}$). Comparison between the reference system (c.f. [5] and section 2.3), and the proposed system is made on manual transcripts and segmentations. However, the named entities detection is automatic and may have some errors. The reference system is described in section 2.3 with two rules using scores $s_l(e_i)$ and $SC_l(e_i)$ (c.f. Equations 2 and 3) and our model is based on belief functions (Equation 9).

As Table 6 shows, in the new model, the error rate in terms of duration ($ErrDur$) is 3 points less than reference system with rule $\mathbf{R}_2$ and 7 points less with rule $\mathbf{R}_1$. Not only the use of belief functions is more easily interpretable, but also it allows to eliminate much errors. Concerning the number of identified speakers, the result is even more obvious:the new system correctly labels much more speakers than the base system, and also improves the reference system. In conclusion, taking account global information on speakers within a speaker turn, and before the decision, allows to significantly improve results both in terms of duration and in terms of numbers of speakers.

Table 6. Comparison between the proposed system and the reference system according to the decision rule on the test corpus of ESTER 1 phase II campaign; **ErrDur**: error rate in duration; **ErrSpk**: error rate in number of speakers.

System	ErrDur	ErrSpk
Reference (rule $\mathbf{R_1}$)	20.6%	20.2%
Reference (rule $\mathbf{R_2}$)	16.6%	19.5%
Proposed (rule $\mathbf{R}$)	**13.7**%	**14.9**%

5 Conclusion

The speaker identification method proposed in this article allows to extract speaker identities from transcriptions. The identification is realized thanks to a semantic classification tree which helps to give the full names found in the transcription to speakers in a recording. In this article, we propose a new system that consistently combines different information about the potential speakers in the form of belief functions. Particularly, the system manages possible conflict of information on the speakers within a speaker turn and takes into account the uncertainty concerning the gender. Experiments have been realized on a French broadcast news and the system have very good performances. Future work will focus on developing solutions to deal with automatic outputs containing errors. Different kind of uncertainty, dues to segmentation error, classification in speakers or to the bad transcription of full names will be taken into account. We will also study the realistic case of open systems when the list of possible speakers is unknown.

References

1. Canseco-Rodriguez, Lamel, L., Gauvain, J.-L.: A comparative study using manual and automatic transcriptions for diarization. In: Automatic Speech Recognition and Understanding, pp. 415–419, San Juan (2005)
2. Deléglise, P., Estève, Y., Meignier, S., Merlin, T.: The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. In: European Conference on Speech Communication and Technology, pp. 1653–1656 (2005)
3. Galliano, S., Geffroy, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier,G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: European Conference on Speech Communication and Technology (2005)
4. Kuhn, R., De Mori, R.: The application of semantic classification trees to natural language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence,17(5), 449-460 (1995)
5. Jousse, V., Petitrenaud, S., Meignier, S., Estève, Y., Jacquin, C.: Automatic named identification of speakers using diarization and ASR systems. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Taipei, pp. 4557–4560 (2009)
6. Mauclair, J., Meignier, S., Estève, Y.: Speaker diarization: about whom the speaker is talking? In: IEEE Odyssey (2006)
7. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
8. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence. 66, 191-234 (1994)
9. S. E. Tranter. Who really spoke when? Finding speaker turns and identities in broadcast news audio. In: In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1013-1016 (2006)

Appendix: Springer-Author Discount

LNCS authors are entitled to a 33.3% discount off all Springer publications. Before placing an order, the author should send an email, giving full details of his or her Springer publication, to `orders-HD-individuals@springer.com` to obtain a so-called token. This token is a number, which must be entered when placing an order via the Internet, in order to obtain the discount.

6 Checklist of Items to be Sent to Volume Editors

Here is a checklist of everything the volume editor requires from you:

- ☐ The final LaTeX source files
- ☐ A final PDF file
- ☐ A copyright form, signed by one author on behalf of all of the authors of the paper.
- ☐ A readme giving the name and email address of the corresponding author.