Combining transcription-based and acoustic-based speaker identifications for broadcast news

Elie El Khoury, Antoine Laurent, Sylvain Meignier, Simon Petitrenaud

COMBINING TRANSCRIPTION-BASED AND ACOUSTIC-BASED SPEAKER IDENTIFICATIONS FOR BROADCAST NEWS

Elie El Khoury[(1,2)], *Antoine Laurent*[(1)], *Sylvain Meignier*[(1)], *Simon Petitrenaud*[(1)]

[(1)]LIUM, Université du Maine – Le Mans, France
[(2)]IDIAP Research Institute – Martigny, Switzerland
firstname.lastname@lium.univ-lemans.fr, ekhoury@idiap.ch

ABSTRACT

In this paper, we consider the issue of speaker identification within audio records of broadcast news. The speaker identity information is extracted from both transcript-based and acoustic-based speaker identification systems. This information is combined in the belief functions framework, which makes coherent the knowledge representation of the problem. The Kuhn-Munkres algorithm is used to optimize the assignment problem of speaker identities and speaker clusters. Experiments carried out on French broadcast news from the French evaluation campaign ESTER show the efficiency of the proposed combination method.

Index Terms— Speaker identification, speaker diarization, belief functions.

1. INTRODUCTION

The speaker identity detection within audio records contains two main steps that are subject to uncertainty and confusion. The first one, often known as "speaker diarization", consists in detecting speakers turns and then clustering those uttered by the same speaker. This problem was studied in our previous works [9]. The next step, often known as "speaker identification", aims to automatically provide - if possible - an identity for each of the resulting clusters. This issue is studied in this paper.

A first approach to identify speakers by their *real* full name (first name and family name) aims to use acoustic *a priori* information for targeted speakers: this requires the availability of training data for each speaker, and thus restricts the targeted speakers to a finite list of speakers.

A second approach to identify speakers by their *real* full name consists in extracting them from the automatic speech recognition system (ASR) [3, 10, 13]. The general principle is to determine if a detected named entity as a "PERSON" refers to a speaker in the document or not. This principle assumes that the names are often pronounced as in broadcast news. This approach has less restriction than the acoustic methods, but it is subject to additional errors due to the use of the ASR system.

In a previous work [10], we proposed a method based on belief functions in order to assign the uttered full names to anonymous speakers. The formalism of belief functions helps to combine the information that belongs to the potential speaker and, contrarily to the probabilities framework, it is particularly suitable to manage the conflicts between speaker identities.

In this paper, the goal is to improve the system proposed in [10] by combining the transcript-based information and the acoustic-based information in order to enhance the system performance.

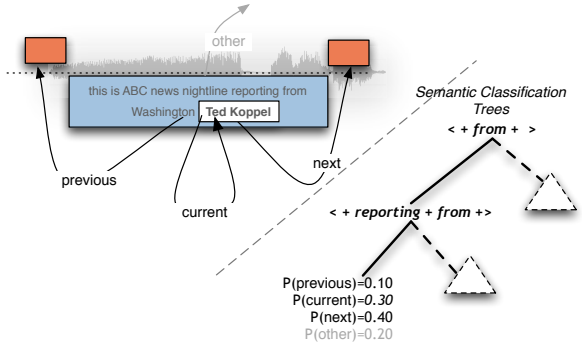

Fig. 1. Illustration of the use of the semantic classification trees

The remainder of the paper is structured as follows: in section 2, we describe the transcript-based speaker identification system. In section 3, we detail our acoustic-based speaker identification system and the fusion between the two systems. In section 4, we describe the experiments and the metrics, and finally present the results obtained on ESTER corpus [6].

2. SPEAKER IDENTIFICATION BASED ON TRANSCRIPTS

The main hypothesis initially proposed in [3] assumes that a detected full name in a speaker turn allows to identify the speaker talking in this turn or in the contiguous turns (previous or next turns). The speaker identification method uses a binary decision tree based on the principle of semantic classification trees (SCT) [7]. A SCT automatically learns lexical rules from full names detected in the training set, 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 of one of the four possible hypotheses: *"current turn"*, *"previous turn"*, *"next turn"* or *"another person"* (see Figure 1). These probabilities are determined during the learning of the tree and reflect the observed cases in the training set.

The used identification method is based on previously transcribed documents in which the named entities are detected. First, the document is cut into segments which are then clustered into anonymous speakers: it is the speaker diarization step. The gender and bandwidth (studio or telephone) of each speaker are also detected. Finally, the segments that are grouped into speaker clusters, are transcribed and the named entities are found.

Here we give some useful notations for the rest of the paper. Let $\Omega = \{\omega_1, \ldots, \omega_I\}$ 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 Ω to a speaker of $\mathcal{C}$. We have to notice that each speaker c_l may be involved in one or several turns. Moreover, several occurrences of full names may be detected in the same turn. For each occurrence of a full name o_j (for $j = 1, \ldots, J$) detected in a speaker turn t_k, let us define by $P(o_j, t_k)$ the probability that o_j is the current speaker. Thus, $P(o_j, t_{k-1})$ and $P(o_j, t_{k+1})$ represent the probabilities that o_j is respectively the speaker of the previous and the next turn given by the SCT.

If the gender of the full name ω_i and the speaker c_l are different, the corresponding occurrence is ignored. The gender of the full name is determined using a linguistic base of first names.

We focus on a turn t_k that has n_k occurrences and that belongs to speaker c_l. Let n_{k+r} be the number of occurrences for the previous turn ($r = -1$) and the next turn ($r = 1$). 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. In [10], we proposed a method based on the mathematical concept of a belief function [11, 12] to assign the full names to anonymous speakers. Each occurrence $o^k_{j,r}$, corresponding to a name ω_i, represents some knowledge concerning the speaker of the turn t_k that can be described by the belief function $m^{jr}_{t_k}$ on Ω, focused on ω_i and Ω:

$$\begin{cases} m^{jr}_{t_k}(\{\omega_i\}) = P(o^k_{j,r}, t_{k-r}) \\ m^{jr}_{t_k}(\Omega) = 1 - P(o^k_{j,r}, t_{k-r}) \end{cases} \quad (1)$$

In this paper, we continue to adopt the point of view proposed by Smets [12]: the Transferable Belief Model (TBM). The representation of the uncertainty is made by the means of a 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. One of the most important operations in the TBM is the procedure for aggregating information in order to combine several belief functions defined in a same frame of discernment [12]. In particular, the combination of two belief functions m_1 and m_2 defined on Ω using the conjunctive binary operator $\cap$, denoted as $m' = m_1 \cap m_2$, is expressed by:

$$\forall A \subseteq \Omega,\ m'(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \quad (2)$$

Repeatedly, we may define the combination of n belief functions $m_1, \ldots, m_n$ on Ω by: $m = m_1 \cap \ldots \cap m_n$.

The first combination step consists in aggregating the whole information in a given speaker turn. In order to ensure associativity and commutativity of the fusion, the combination of the $n_{k-1}+n_k+n_{k+1}$ belief functions defined on the turn t_k by Equation 1 is made with the conjunctive *non normalized* rule (Equation 2). We obtain a belief function m_{t_k} that represents a more synthetic knowledge of speaker identity provided in turn t_k. It is defined as:

$$m_{t_k} = \bigcap_{r=-1}^{1} \bigcap_{j=1}^{n_{k+r}} m^{jr}_{t_k} \quad (3)$$

The second combination step consists in aggregating the results obtained for each speaker turn of a given speaker c_l. Therefore we combine all the belief functions corresponding to the speaker turns t_k of this speaker with the same conjunctive rule and we obtain a global belief function M_l which represents the state of belief concerning speaker c_l for the whole audio record:

$$M_l = \bigcap_{t_k \in c_l} m_{t_k} \quad (4)$$

In a final decision step that enables the assignment of a full name to an unknown speaker, we propose to use the famous "Kuhn-Munkres algorithm" [2] to optimize the assignment problem. We define the cost function according to the masses obtained in Equation 4:

$$C(\omega_1, \ldots, \omega_I) = \sum_{l=1}^{L} M_l(\omega_i) \quad (5)$$

Then the cost function is maximized by the Munkres algorithm by the use of this global criterion. The decision process is simple and unified thanks to the use of belief functions.

An example is given in Table 1: three different speakers have their maximum mass M_l corresponding to the same full name "Jacques Derrida" (see Equation 4), but according to the Munkres algorithm, this full name is finally assigned to c_3. (see Table 1).

Table 1. Decision with the Munkres algorithm (decision in bold, belief masses $M_l(e_i)$ in parentheses).

Cluster	Full name (in decreasing order)
c_1	J. Derrida (0.79), **N. Demorand** (0.11)
c_2	J. Derrida (0.51), **A. Adler** (0.25)
c_3	**J. Derrida** (0.98), O. Duhamel (0.001)
c_4	**O. Duhamel** (0.78), J. Derrida (0.08), A. Adler (0.02)
c_5	**Marc Kravetz** (0.84) Jacques Derrida (0.02)

3. ADDING ACOUSTIC-BASED INFORMATION

3.1. Speaker identification based on GMM

Our acoustic speaker identification system is based on the well-known UBM/GMM approach developed in the ALIZE toolkit[1]. First, a pre-processing concerning training, development and test sets is necessary. It is composed of three main steps:

1. **Feature extraction:** 19 Mel Frequency Cepstral Coefficients (MFCC) and their first order derivatives are computed every 10 ms. The energy and delta-energy are also used to construct acoustic vectors of 40 elements.
2. **Non-speech removal:** after normalizing the energy coefficient, energy detector based on three components GMM classifier is used to separate speech from non-speech.
3. **Feature normalization:** parameter vectors that correspond to speech are normalized to fit a zero mean and unit variance distribution.

Then on the training set, the universal background model (UBM) and the models of targeted speakers are generated:

[1] http://alize.univ-avignon.fr/

1. **World model training:** after discarding the energy coefficients from the acoustics vectors, a general UBM that represents all the targeted speakers is estimated using Expectation-Maximization (EM) algorithm. It is the concatenation of 4 UBMs (Male-Studio, Male-Telephone, Female-Studio, Female-Telephone) each composed of 128 Gaussian distributions.
2. **Speaker model training:** the UBM parameters (means and variances) are adapted to the speaker training data using Maximum *A Posteriori* (MAP) estimation.

The training phase is followed by the computation of the scores between the anonymous speakers and the targeted speakers. This concerns both development and test sets:

1. **Score computation:** for each of the anonymous speakers (obtained at the end of the speaker diarization task), a list of scores are computed, each is equal to the log likelihood ratio (LLR) for this speaker given the UBM and the targeted speaker model. To quickly compute these scores, only the 10 "top Gaussian distributions" of the models are considered.
2. **Score normalization:** to cope with the score variability between anonymous speakers, the Test normalization technique (T-Norm) [4] is applied.

On the development set, we can determine a cut-off threshold ($T_{cut-off} = 4.1$) that removes the lowest scores. $T_{cut-off}$ is then used for the test set.

3.2. Fusion of transcript-based and acoustic-based information

The simplest way to combine the two systems described in previous sections is to transform the score provided by the acoustic-based system into a belief function. On the development set, we compute the mean (μ) and the standard deviation (σ) that correspond to the scores distribution of the "True speakers" (i.e. when the anonymous speaker matches the targeted speaker). On the test set, μ and σ are used to compute the belief function: in this work, we use the Cumulative Distribution Function value (F_i) of the score where F_i corresponds to the name ω_i. Thus, we obtain a belief function M_l^{Gmm} for each anonymous speaker c_l:

$$\begin{cases} M_l^{Gmm}(\{\omega_i\}) = F_i \\ M_l^{Gmm}(\Omega) = 1 - F_i \end{cases} \quad (6)$$

This belief function is combined with the one obtained on the transcript-based system (see Equation 4) according to the conjunctive rule:

$$M_l^{Fus} = M_l^{Gmm} \cap M_l \quad (7)$$

Finally, we apply the Munkres algorithm with this new set of belief functions in order to identify the anonymous speakers.

4. EVALUATION OF THE PROPOSED SYSTEM

4.1. Data description, tools and metrics

The experiments are realized on the corpus of the French evaluation campaign ESTER 1 [6]. The data were recorded from 6 French-speaking radios. They last from 10 to 60 minutes. This corpus is divided into 3 sets that are used for the training, the system development and the evaluation. For each of those sets, the total duration, the number of speaker turns and the number of full name occurrences are reported in table 2. During experiments, the transcript-based system is supposed to know all the full names of the speakers. This list is composed of 1008 full names extracted from the corpus. Moreover, there are 349 speakers (237 males and 112 females) for whom 2 minutes of acoustic records are available.

Table 2. Corpus description

set	duration	speaker turns	full name
Train	76 h	7416	11292
Dev.	30 h	2931	4533
Test	10 h	1082	1541

The diarization system [9] is composed of an acoustic BIC-based segmentation followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian distribution. Viterbi decoding is used to adjust the segment boundaries using GMMs for each cluster. Music and jingle regions are removed using also Viterbi decoding. Gender and bandwidth are then detected. This system, completed by a CLR-based clustering phase, obtained the best results during the ESTER 2 campaign and the diarization error rate (DER) is 7.27% on the ESTER 1 test set.

The ASR system is based on two-pass decoding. The best hypotheses generated by pass 1 permit to compute a CMLLR transformation for each speaker. Decoding pass 2 uses Speaker Adaptive Training (SAT), Minimum Phone Error (MPE) acoustic models and the CMLLR transformations. The full LIUM ASR system (see [5] for more details) was not used because language models (LM) are based on class models in our experiments. Only the decoder of the passes 1 and 2 implement the class LM. The LM contains 121K words and one class representing the name of the speakers. This class contains the 1008 full names. The word error rate (WER) is 13.6% for the test set.

The named entity detector employed is the LIA_NE [1] based on a mixed approach with generative models (HMM) and discriminative models (Conditional Random Field). This system obtained the best results at the name entity recognition task on ASR transcripts during ESTER 2 evaluation campaign. The slot error rate is 23.9% on manual transcripts and 51.6% on LIUM ASR transcripts (official results of the campaign). For the "PERSON" entity, the slot error rate is 11% using manual transcripts and 59% using LIUM ASR transcripts.

The proposed system is evaluated comparing the generated hypothesis and the reference. This comparison highlights 3 error rates computed in terms of duration:

- Substitution error (Sub): system hypothesis differs from the one found in the reference.
- Deletion error (Del): no identity is proposed by the system but the speaker is identified in the reference.
- Insertion error (Ins): an identity is proposed by the system but the speaker is unknown in the reference.

The total error rate, defined in [13], is the sum of those 3 errors:

$$Err = Sub + Ins + Del \quad (8)$$

4.2. Results

The various parameters are tuned using the development set. The proposed systems are evaluated using either manual or automatic segmentations and transcripts. In the two cases, the named entities are automatically detected using [1].

Table 3. Results on the test set.

system	Sub	Del	Ins	Err
Using reference segmentations and transcripts				
Transcript-based system	3.57	6.29	**0.20**	10.06
GMM-based system	1.82	58.05	2.20	62.07
Transcript+GMM system	**1.31**	**0.94**	2.40	**4.65**
Using automatic segmentations and transcripts				
Transcript-based system	14.25	25.69	**1.24**	41.17
GMM-based system	**2.67**	57.15	3.13	62.95
Transcript+GMM system	13.85	**14.69**	4.17	**32.72**

The progress done on all elementary tasks (speaker diarization, speech recognition and the use of belief functions) decreases the speaker identification error rate from 75.15% [8] to 60.59%.

Moreover, the use of the **LM class transcription strategy** (13.6% of WER) provides an additional gain of 19.42 points with an error rate equal to 41.17% (see table 3). We have to notice that the classical LIUM transcription system with 5 passes that does not use LM class is slightly better (11.4% of WER).

Table 3 also shows that the error rate using fully automatic system (41.17%) is 4 times higher than the results using the reference (10.06%). In an additional experiment, we have found that near half of the errors are due to the diarization errors: the diarization system was developed to minimize the DER, but the small speaker turns ($< 2s$) are often missed and generate bad SCT detection.

The GMM-based system has a high error rate of 62.95% using the automatic segmentations and transcripts but 58.05% of the errors are coming from deletion because the GMM models of the undetected speakers do not exist.

Table 4 shows contrastive results, where the scoring is limited to targeted speaker for whom the audio records are available. The error rate of the GMM-based system is 14.36% with 8.22% of deletion mostly due to the cut-off on LLR scores.

In both tables 3 and 4, the GMM system is less affected by the diarization errors: only around 1 point is lost using the automatic diarization.

In all cases, **combining both transcript-based and acoustic-based** information gives the best results. The error is 32.72% on the automatic segmentations and transcripts (a gain of 8.45 points) and 4.65% on the references (a gain of 5.41 points) as seen in table 3. Results on the restricted list of GMM targeted speakers lead to the same conclusion (table 4).

5. CONCLUSION

The speaker identification method proposed in this paper allows to extract speaker identities from acoustic records of broadcast news. Other than the improvements of the transcript-based system, we propose a new system that consistently combines acoustic-based and transcript-based information of the potential speakers in the framework of belief functions. Particularly, the system manages possible conflict of information on the speakers. Experiments done on French broadcast news show the efficiency of the proposed system with a speaker identification error rate of 4.65% on manual segmentation and transcripts and 32.72% when the system is fully automatic. Future work will focus on improving the diarization task in order to fit the problem of speaker identification. Moreover, visual information will be added to the framework of belief functions in order to deal with the audiovisual speaker identification on TV shows.

Table 4. Results limited to the GMM targeted speakers list.

system	Sub	Del	Ins	Err
Using reference segmentations and transcripts				
Transcript-based system	1.86	6.47	**1.06**	9.39
GMM-based system	0.50	8.94	3.86	13.31
Transcript+GMM system	**0.26**	**0.88**	2.83	**3.98**
Using automatic segmentations and transcripts				
Transcript-based system	3.62	13.14	**4.82**	21.58
GMM-based system	**1.17**	8.22	4.97	14.36
Transcript+GMM system	2.44	**2.92**	8.65	**14.01**

6. ACKNOWLEDGMENT

The research leading to these results is part of the ANR-SODA project that is funded by the French research agency.

7. REFERENCES

[1] F. Béchet and E. Charton. "Unsupervised knowledge acquisition for extracting named entities from speech". IEEE ICASSP, USA, 2010.

[2] R. E. Burkard, M. Dell'Amico, and S. Martello. "Assignment Problems". *SIAM*, 2009.

[3] L. Canseco, L. Lamel, and J.-L. Gauvain. "A comparative study using manual and automatic transcriptions for diarization". *Automatic Speech Recognition and Understanding*, 2005.

[4] M. Carey, R. Auckenthaler and H. Lloyd-Thomas. "Score normalization for text-independent speaker verification systems". *Digital Signal Processing*, 2000.

[5] P. Deléglise, Y. Estève , S. Meignier and T. Merlin. "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?". Interspeech, United Kingdom, 2009.

[6] S. Galliano, E. Geffroy, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news". *European Conference on Speech Communication and Technology*, 2005.

[7] R. Kuhn, and R. De Mori. "The application of semantic classification trees to natural language understanding". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[8] V. Jousse, S. Petitrenaud, S. Meignier, Y. Estève, and C. Jacquin. "Automatic named identification of speakers using diarization and ASR systems". *IEEE ICASSP*, Taiwan, 2009.

[9] S. Meignier and T. Merlin. "LIUM SpkDiarization: an open source toolkit for diarization". CMU SPUD Workshop, USA, 2010.

[10] S. Petitrenaud, V. Jousse, S. Meignier and Y. Estève. "Speaker identification using belief functions". Information Processing and Management of Uncertainty (IPMU), Germany, 2010.

[11] G. Shafer. "A Mathematical Theory of Evidence". *Princeton University Press*, 1976.

[12] P. Smets, and R. Kennes. "The transferable belief model". *Artificial Intelligence*, 1994.

[13] S. E. Tranter. "Who really spoke when? Finding speaker turns and identities in broadcast news audio". *IEEE ICASSP*, Taiwan, 2009.