AUTOMATIC NAMED IDENTIFICATION OF SPEAKERS USING DIARIZATION AND ASR SYSTEMS

Vincent Jousse, Simon Petit-Renaud, Sylvain Meignier, Yannick Estève, Christine Jacquin

LIUM (Le Mans), France - LINA (Nantes), France

ABSTRACT

In this paper, we consider the extraction of speaker identity from audio records of broadcast news without *a priori* acoustic information about speakers. Using an automatic speech recognition system and an automatic speaker diarization system, we present improvements for a method which allows to extract speaker identities from automatic transcripts and to assign them to speech segments.

Experiments are carried out on French broadcast news records from the ESTER 1 evaluation campaign. Experimental results using outputs of automatic speech recognition and automatic diarization are presented.

Index Terms— Named identification, Speaker diarization, Automatic transcription

1. INTRODUCTION

Very large collections of audio/video documents are now available online. They have to be indexed to allow later retrieval of recorded information. Automatic rich transcription can be used at a reasonable cost when specific meta-data is wanted such as the main topic, keywords or the name of the speakers. In this paper, we focus on speaker identification by name.

The first step to automatically get rich transcription consists in detecting homogeneous audio segments which contains the voice of only one speaker; the resulting segments are then clustered by speaker. This step is called diarization. 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]. 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. To identify speaker by name, it is possible to use acoustic *a priori* information for targeted speakers [3]: this implies the availability of training data for each speaker, and the restriction of the targeted speakers to a finite list of known speakers. This makes such systems difficult to manage and to deploy.

A promising approach to identify speaker by name consists in using the outputs provided by an automatic speech recognition (ASR) system. A first system, using manual rules [4], was proposed in 2005. In 2006, we have proposed an automatic approach exploiting ASR transcriptions and the outputs of a diarization system [5]. This approach is based on the use of semantic classification trees (SCT). A close approach was also proposed the same year using *n-grams* [6]. In 2007, an experimental comparison of these two propositions on French broadcast news recordings showed that the SCT-based approach is more robust and significantly more efficient than the *n-gram* one on automatic transcriptions [7]. The same year, [8] proposed an approach using a conditional maximum entropy model for the same task and compared it to the *n-gram* approach. But experiments were made only on manual transcription, manual diarization and manual named entity detection.

In this paper, we present improvements concerning the SCT-based method [5, 7]. In this approach, decisions are taken at two levels: at the segment level (*local decisions*), and at the audio file one (*global decisions*). The main improvement concerns the algorithm used to take global decision, but other new features have been added, as the combination of acoustic and linguistic information.

Last, this paper presents results of an entirely automatic system, using automatic diarization, automatic speech recognition, and automatic named entity detection.

Section 2 presents new features improving the SCT-based method and section 3 describes experiments. Results are presented in section 4.

2. IMPROVEMENTS

2.1. Baseline system

Our system to identify speakers by name is based on the use of SCT and it is presented in [5] and [7]. A SCT is used for each occurrence of full name detected in the transcripts. The SCT allows to associate a tag to each occurrence of full name. This is called a *local decision*. This tag indicates if the full name corresponds to the speaker of the *previous* audio segment, to

This research was supported by the Région des Pays de la Loire and by the ANR (French National Research Agency) under contract number ANR-06-MDCA-006.

the *current* one, to the *next* one, or to *another* person. In fact, the SCT gives a probability for each one of these tags but only the tag associated to the maximal probability is considered.

The diarization system used to segment the audio file has clustered segments by speaker. Using this information and summing the scores of full names linked to each cluster using tags provided by the SCT allows to associate a full name to a cluster. This is the *global decision*.

In this study, *a priori* information is added to improve the *global decision process*.

2.2. Gender usage: combining acoustic and linguistic features

Thanks to the acoustic segmentation and classification processes, the system is informed about the gender of each speaker in the record. It seems to be obvious not to attribute a masculine (respectively feminine) first name to a feminine speaker (respectively masculine). It is one of the improvement added to the system. The system uses a database[1] composed of about 20000 first names each followed by the number of times they were labelled with each gender (masculine or feminine) since 1900. The gender associated to a name is the more frequent one (more than 75% since 1990). If it cannot be associate to a gender, it is considered as unknown. This list is used when the system has to decide which candidate to attribute to an anonymous cluster. It will eliminate all the possible first name/last name couples which don't match (according to the gender of the first name) the gender of the cluster. If a possible speaker has a first name which is undetermined, he is still considered. The entire decision process is explained in 2.3.

2.3. New decision process

2.3.1. Notations

Let $\mathcal{E} = \{e_1, \ldots, e_I\}$ denotes the set of full names hypotheses to assign to a cluster, $\mathcal{O} = \{o_1, \ldots, o_J\}$ the occurrences of full names detected in the transcripts and $\mathcal{C} = \{c_1, \ldots, c_K\}$ the clusters of anonymous speakers to be labeled. For each $j = 1, \ldots, J$, let $P(o_j, r)$ be the probability that o_j is the previous ($r = 1$), the current ($r = 2$) or the next ($r = 3$) speaker respectively and $h_r(o_j)$ be the previous speaker ($r = 1$), the current speaker ($r = 2$) and the next speaker ($r = 3$) when o_j is named.

2.3.2. Computation of scores

In order to avoid useless or noisy information, probabilities $P(o_j, r)$ are filtered. Thresholds α_r ($r = 1, 2, 3$), from which probabilities $P(., r)$ are taken into account, are learned from the training set. Then if $P(o_j, r)$ is under threshold α_r, then $P(o_j, r)$ is set to 0. This precaution prevents from accumulative little errors.

A second filter is made by the comparison of genders: if the gender of the full name e_i and the cluster c_k are different, the probabilities are not taken into account in the computation of the scores. Let $g(e_i)$ and $g(c_k)$ be the gender (female, male or unknown) of a full name e_i (or a cluster c_k), then: $(o_j = e_i,\ h_r(o_j) = c_k$ and $g(e_i) \neq g(c_k)) \Rightarrow P(o_j, r) = 0$.

For the assignment to a given c_k, we compute a score for each full name e_i, denoted as $s_k(e_i)$, as a simple sum of the filtered probabilities:

$$s_k(e_i) = \sum_{\{(o_j,r)|o_j=e_i,\ h_r(o_j)=c_k\}} P(o_j, r) \qquad (1)$$

Let $\mathcal{D} = \{c_k \in \mathcal{C} |\ \forall e_i \in \mathcal{E}, s_k(e_i) = 0\}$ be the set of unlabeled speaker.

2.3.3. Decision process

The aim is now to assign a full name to each cluster. Let $f : \mathcal{C} \rightarrow \mathcal{E}$ be the assignment function of full names to clusters.

The first step naturally consists in choosing the full name which has the maximum score for a given cluster (if there is at least one non-null score):

$$\begin{cases} \forall c_k \in \mathcal{C} \setminus \mathcal{D}, & e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{s}_k(e_i) \Rightarrow f(c_k) = e_i^* \\ \forall c_k \in \mathcal{D}, & f(c_k) = \text{Anonymous} \end{cases} \qquad (2)$$

An issue is that the same full name e_i may be assigned to several clusters. Since the segmentation in clusters is assumed correct, we propose to reorganize the sharing of full names among clusters.

Several strategies may be considered to rank the competing clusters c_k for a given full name e_i. Taking the maximum score $s_k(e_i)$ seems to be the more natural way, but it seems also judicious to use the relative scores of e_i among all the possible candidates, i.e. $sn_k(e_i) = \dfrac{s_k(e_i)}{\sum_{q=1}^{I} s_k(e_q)}$. Indeed, using the normalized scores is a better way to compare competitive full names. However, it may cause problems, particularly in the frequent case when a false full name with a weak score but with no competition may be affected to a cluster.

In order to avoid these problems, we propose to adopt a compromise as the product of normalized and non-normalized scores:

$$\mathrm{SC}_k(e_i) = \frac{s_k^2(e_i)}{\sum_{q=1}^{I} s_k(e_q)} \quad \text{if } c_k \notin \mathcal{D} \qquad (3)$$

and $\mathrm{SC}_k(e_i) = 0$ if $c_k \in \mathcal{D}$.

A concrete example is given in Table 1. The full name "Jacques Derrida" has been assigned to three different clusters. In this example, c_{13} has the best score and should be finally labeled as "Jacques Derrida" but the score represents

[1] Collected from various web sources

only 39% of the total scores among all the possible candidates for c_{13}, whereas the score for c_{15} is 79%. Finally "Jacques Derrida" is assigned to c_{15} and the clusters c_{13} and c_{14} will be labeled with another full name.

Table 1. Example of an initial multiple assignment

Cluster	full name e_i^*	$s_k(e_i)$	$sn_k(e_i^*)$	$\text{SC}_k(e_i)$
c_{13}	Jacques Derrida	**8.58**	0.39	3.36
c_{14}	Jacques Derrida	1.67	0.65	1.09
c_{15}	Jacques Derrida	4.94	**0.79**	**3.88**

The second step consists in reassigning full names to speakers in case of multiple assignment.

The decision process presented in [7] consists in selecting the full name e_i with the maximum score $s_k(e_i)$ for the cluster c_k.

With the new decision process, all the possible full names are *a priori* taken into account and sorted according to their score $\text{SC}_k(e_i)$. First, the full name with maximum score is chosen and if several clusters are associated with a same e_i^* (*i. e.* $|f^{-1}(e_i^*)| > 1$), then this full name will be assigned to the cluster whose score $SC_k(e_i^*)$ is maximum. Then, all chosen full names are deleted from the list of clusters that are not yet assigned in this first round. In the second round, remaining full names are examined in the same way for the remaining clusters and so on, until all clusters 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 (decision in bold type, scores in parenthesis).

Cluster	full name e_i^* (1st choice)	2nd choice
c_{13}	Jacques Derrida (3.36)	**Nicolas Demorand** (0.99)
c_{14}	Jacques Derrida (1.09)	**Alexandre Adler** (0.30)
c_{15}	**Jacques Derrida** (3.88)	Olivier Duhamel (0.02)
c_{16}	**Olivier Duhamel** (0.93)	Jacques Derrida (0.14)

3. EXPERIMENTS

3.1. Data description

The methods are trained and developed with data from the ESTER 1 (2003-2005) evaluation campaign [9].

The data were recorded from six radios : *France Inter, France Info, RFI, RTM, France Culture* and *Radio Classique*. They are divided in 3 corpora: the training corpus of 81h (150 shows), the development corpus of 12.5h (26 shows) and the test corpus of 10h (18 shows). 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.

3.2. Automatic system

3.2.1. Diarization

The LIUM diarization system was developed for the transcription task of the ESTER evaluation campaign. It is composed of an acoustic BIC segmentation, which is followed by a BIC hierarchical clustering. Each cluster is modelled with a full covariance Gaussian. Viterbi decoding is performed to adjust the segment boundaries.

Music and jingle regions are removed using Viterbi decoding. The decoding uses 8 GMMs corresponding to 2 silences (wide and narrow band), 3 wide band speeches (clean, over noise or over music), 1 narrow band speech, 1 music and 1 jingle. The GMMs contain 64 diagonal Gaussians and are trained by EM-ML on ESTER data.

Speaker clustering is filtered, only speech areas are kept. At the end, a CLR hierarchical clustering is computed over the filtered speaker clusters.

3.2.2. Automatic Speech Recognition system

Experiments on speech recognition were carried out by using the LIUM ASR system, based on the CMU Sphinx 3.x decoder, described in [10]. It is a three-pass system: the first pass uses a trigram language model and generic acoustic models (one for each of the four gender/band conditions — female/male + studio/telephone), the second pass uses the best hypothesis of the first pass to adapt the acoustic models using SAT and CMLLR, and the last pass consists in rescoring with a quadrigram language model a word-graph generated during the second pass. This system ranked second in the French ASR evaluation campaign ESTER, and was the best open source system [9].

3.3. Scoring

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.

Precision, recall and error rate are defined as follows:

$$P = \frac{C_1}{C_1 + S + I} \;\; ; \;\; R = \frac{C_1}{C_1 + S + D} \tag{4}$$

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1} \;\; ; \tag{5}$$

System	Recall	Precision	Err	Err Spk
Baseline	70.70%	92.59%	26.64%	37.40%
New system	83.16%	89.72%	16.66%	19.5%

Table 3. Results comparison on test corpus

Err*: Error rate (in duration)*
Err Spk *: Speaker error rate (in number of speakers)*

There are two ways of assessing the results. In all the previous papers dealing with named speaker identification, the results were presented in terms of duration [4, 5, 6, 7, 8]. That is to say that if a system is able to correctly name a speaker who speaks 90% of the time and miss the other six speakers who speak 10% of the time, it will have very good results (90% of recall and 90% of precision). However in this case it has found only one speaker of the seven speakers in the show. So the system has an error rate of 87,5% considering the number of speakers found (one of seven).

In our results, we will use those two ways of scoring: in duration (naming a speaker who speaks a lot is important) and in number of speakers (the most important thing is to name as much as possible speakers).

4. RESULTS

The results are presented in tables 4 and 3. The named entity detection is always performed automatically. Compared to the baseline system on manual diarization and transcription, the new system has a worse precision measure (about 3% less) but a better recall (more than 12% better) in term of duration. Focusing on the number of speakers found, the new system is able to find about twice as many speakers comparing to the old system. Indeed, the error rate considering the number of speakers found is about 20% for the new system against almost 40% for the old.

When dealing with automatic outputs, there is a big performance drop. Indeed, using automatic diarization (with manual transcription) or automatic transcription (with manual diarization) gives a speaker error rate of 70%. With a fully automatic system, the error rate is even worse: about 85%.

The main problem with automatic transcription is the error made in the transcription of the names. It affects both recall and precision: some names are not well transcribed and not detected by the named entity detection system. When they are detected they are mostly not well spelled: the precision of the system decreases. Specific solutions have to be developed to deal with automatic outputs.

5. CONCLUSION

Improvements concerning the SCT-based method were presented in this paper. With the new decision process and the combination of heterogenous information, the system is able to find twice as many speakers than the former one. Moreover, those improvements can be integrated in other methods like the n-gram one. In this study, experiments on speaker identification by name are made in an entirely automatic way: all the data are provided by automatic systems (ASR and diarization), and an automatic named entity detection system was used. It is the first time that such experimental results are presented about this task. Future work will focus on developing solutions to deal with automatic outputs: improving automatic classification, transcription of first names, working on n-best ASR outputs and using more acoustic features.

Trans.	Diar.	R	P	Err	Err Spk
M	**M**	83.16%	89.72%	16.66%	19.5%
M	**A**	38.02%	58.19%	58.25%	71.19%
A	**M**	30.98%	58.3%	62.8%	69.96%
A	**A**	18.36%	42.08%	75.15	84.77

Table 4. New system results with automatic outputs

Trans.*: **M**anual or **A**utomatic transcription.*
Diar.*: manual or automatic speaker diarization.*
R, P*: recall and precision (in duration).*

6. 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 *ICSLP*, Korea, 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*, vol. 4, pp. 430–451, 2004.

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

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

[6] S. E. Tranter, "Who really spoke when? Finding speaker turns and identities in broadcast news audio," in *ICASSP*, Toulouse, France, May 2006, vol. 1, pp. 1013–1016.

[7] Y. Estève, Sylvain Meignier, and Julie Mauclair, "Extracting true speaker identities from transcriptions," in *Interspeech*, Antwerp, Belgium, August 2007.

[8] M. Chengyuan, Patrick Nguyen, and Milind Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *ICASSP*, Honolulu, HI, USA, April 2007.

[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," in *Interspeech 2005*, Lisbon, Portugal, September 2005.

[10] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news," Lisbon, Portugal, September 2005.