Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both ?

Johann Poignant[1]*, Laurent Besacier*[1]*,*
Viet Bac Le[2]*, Sophie Rosset*[3]*, Georges Quénot*[1]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

[2]Vocapia Research, 28 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France

[3]Univ Paris-Sud, LIMSI-CNRS, Spoken Language Processing Group, BP 133, 91403, Orsay, France

[1]first.lastname@imag.fr, [2]levb@vocapia.com,[3]first.lastname@limsi.fr

Abstract

Persons identification in video from TV broadcast is a valuable tool for indexing them. However, the use of biometric models is not a very sustainable option without a priori knowledge of people present in the videos. The pronounced names (PN) or written names (WN) on the screen can provide hypotheses names for speakers. We propose an experimental comparison of the potential of these two modalities (names pronounced or written) to extract the true names of the speakers. The names pronounced offer many instances of citation but transcription and named-entity detection errors halved the potential of this modality. On the contrary, the written names detection benefits of the video quality improvement and is nowadays rather robust and efficient to name speakers. Oracle experiments presented for the mapping between written names and speakers also show the complementarity of both PN and WN modalities.

Index Terms: Speaker identification, OCR, ASR

1. Introduction

Nowadays, with the growing number of audio-visual content available, the automatic identification of people appears as very useful for searching and browsing in this type of data. Such person identification may for instance be based on speaker recognition technology. However, training biometric models of speakers requires costly manual annotations of video contents.

As we can not consider the manual annotation of each new video source as a viable option, an interesting alternative is the use of unsupervised approaches for naming people in multimedia documents. To this end, we can automatically classify each speech turn with an anonymous label (i.e. speakers clustering or diarization) and use others sources of information that provide the real person names for at least some of the clusters. When dealing with TV broadcast, at least two different modalities can provide the real names of the persons speaking: (i) the names extracted from the speech transcript (ASR output) and (ii) the names written on the screen by the show to introduce a person (name in OCR output written in a title block[1]).

In Figure 1, we can see an example from a TV news show including an anchor, a journalist and a person interviewed. In this example, the arrows represent the citation links from written names and pronounced names to a person appearing/talking in the show. Indeed, in this example, there is a correlation between names pronounced or written and the audio-visual presence of this person in the adjacent speech turns or shots.

Naming people in television programs using automated systems can allow to further address several tasks:
Automatic annotation: it can help/complement/replace the manual annotation. This task is recall and precision oriented.
Creating models: The automatic extraction of audio segments can be used to build speakers models. Such a task is driven by the accuracy of the speaker models that must be as pure as possible while having enough signal to generate them.
Information retrieval: The answer to a query propose several video segments where a person is present. This task is recall oriented and should be able to hande the maximum number of persons (even those for which no a priori model is available).

This paper focuses on naming speakers in TV broadcast. The pronounced names and written names provide both relevant information to answer to these questions. Previous works mainly used pronounced names. The names written on the screen were seldom used due to the poor image quality which lead to low performance text detection and transcription systems. But the evolution of video quality available brings us to reassess the use of this modality. We therefore propose a comparative study of the potential of pronounced names (obtained via ASR) and written names (obtained via OCR) to identify/name a speaking person in a TV broadcast.

This article begins with an overview of the literature on naming people in broadcast radios and videos. More particularly, we focus on the methods for the extraction of hypothesis names (clustering methods and association name-person being outside the scope of this literature review). Then, we continue with a presentation of the REPERE corpus on which we experimented. Next, we compare the quality of the extraction of pronounced names (ASR) and written names (OCR) using automatic systems. Finally, we evaluate both modalities for an unsupervised speaker detection task on TV broadcast. This is done with an oracle mapping (adjacent speech turns) between person-name whatever the time stamp of the name citation.

2. State-of-the-art

Previous works concerning the unsupervised naming of people in television or radio, use essentially the same framework:

- Persons clustering (diarization).
- Hypothesis names extraction for each person.
- Hypothesis names/persons mapping (or association).

This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

[1]Title block: spatial position used by the show to write a name in order to introduce the corresponding person.

Figure 1: Pronounced names and written names in a TV broadcast video

Thereafter, we will focus on methods of extracting hypothesis names for each person. Pronounced names (PN) are mostly used in the state of the art due to the poor quality of the written names (WN) transcription. In the papers we reviewed, three steps, either manual or automatic, were generally used:

- Detection and transcription of the speech or written text on the screen.
- Person names detection in the transcription.
- Mapping of each hypothesis name to a speaker.

The first works were proposed by *Canseco et al.* in [1] and [2]. The authors use linguistic patterns set manually in order to determine to which a pronounced name refers: the current speaker ("Hello, I am Joie Chen"), following ("This is Candy Crowley") or previous ("thank you Candy Crowley"). *Tranter et al.* [3] replace manual rules by a learning phase of n-grams sequences with associated probabilities. *Mauclair et al.* [4] use a semantic classification tree trained to associate a pronounced name to a speaker. *Estève et al.* [5] compare these two techniques. They conclude that the semantic classification trees are less sensitive than the sequences of n-grams when using automatic speech transcriptions. *Jousse et al.* [6] improve the using of the semantic classification trees with a local decision (affiliate a name to a nearby speech turn) and a global decision (propagation names into speaker clusters). They also show a performance degradation between 19.5% and 70% relative (speakers identification error rate) when using automatic speech transcriptions instead of manual transcriptions. More recently, in [7] we proposed three propagation methods to map written names to speakers clusters. These unsupervised methods, inherently multi-modal, get much better performance than a mono-modal supervised solution. We have shown that automatic mapping of written names to speakers clusters lead to an accuracy of 98.9% when the diarization is considered as perfect.

The use of automatically extracted pronounced names (PN) faces several challenges: (i) transcription errors. (ii) errors in the person names detection: missing name parts, false detection or adding/removing words (name = "Here John Chan"). (iii) Mapping (affiliation) errors: to which speaker associate a name? The current speaker, the next one, the previous one ?

The use of automatically extracted written names (WN) faces the same difficulties: (a) transcription errors: the increase of the video quality reduces these errors. (b) errors in the person names detection: each show uses a template with specific locations to write texts. The difficulty lies in the detection of spatial positions of title blocks. (c) Mapping (affiliation) errors: usually a name is written on the screen while person is talking.

3. REPERE Corpus

Our comparison of these two modalities for unsupervised speaker detection will be based on the REPERE corpus [8]. This corpus is composed of 7 different shows (news, talks, debates) recorded on two French TV channels. The quality of these recordings (720*576, mpeg2) allows us to use the texts written on the screen. For the experiments, we use the training part (phase 1) of this corpus (58 hours of raw video, 24 hours annotated). For the REPERE challenge these videos were partially annotated on UEM segments[2]. On these segments, the speech transcription was completely manually annotated. 555 speakers were named while 255 others were not. These unknown speakers correspond to 25 minutes of speech time among the 1440 minutes annotated. In this paper, we are interested only in speakers who have been identified during the annotation. The written texts are partially annotated (in average one image every ten seconds, i.e. so manual annotation can miss names). These texts are cut out and transcribed, the names of people have been labeled. Raw videos are longer than the annotated segments: advertisings and extra programs may be contained in the raw video. This additional (un-annotated) signal can help to extract more names, to have more occurrences of a name, or to find rarely pronounced or rarely written names (which is actually the case for anchors).

4. Automatic names extraction

4.1. Written names (WN)

To detect the names written on the screen used to introduce a person, a detection and transcription system is needed. For this task we used LOOV [9] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos. We obtained a character error rate (CER) of 4.6% for any type of text and 2.6% for names written on the screen to introduce a person.
From the transcriptions, we use a simple technique for detecting the spatial positions of title blocks. This technique compares each transcript with a list of famous names (list extracted from Wikipedia, 175k names). Whenever a transcription corresponds to a famous name, we add its spatial position in a list. With the repeating positions in this list we find the spatial positions of title blocks used to introduce a person. However, these text boxes detected do not always contain a name. A simple filtering based on some linguistic rules allows us to filter false positives. Transcription errors are corrected using our Wikipedia list when the edit distance is small (207 corrections with 4 errors).

[2]UEM: Evaluation unpartitioned Map

4.2. Pronounced names (PN)

A state-of-the-art off-the-shelf Speech-To-Text system for French [10] was used to transcribe the audio data without specific model adaptation on our corpus. The recognizer uses the same basic statistical modeling techniques and decoding strategy as in the LIMSI English BN system [11]. Prior to transcription, segmentation and clustering [12] are performed. Word decoding is carried out in a 1xRT single decoding pass. Each decoding pass produces a word lattice with cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities (35-phone set, 65k word vocabulary). This system obtained a word error rate of 16.87% (on around 36k words) during the first evaluation campaign of the REPERE challenge. For named-entity detection, we trained specific independent CRF models on the Quaero data. These models used the same features as those presented in [13]: (1) Standard features like word prefixes and suffixes. (2) Morpho-syntactic features extracted as in [15]. (3) Features extracted from a multilevel analyzer used in the LIMSI question-answering systems [16].

4.3. Comparison of WN and PN systems quality

The use of LOOV pipelined with our written names detection technique allows us to obtain 97.7% of names (see Table 1), with a precision of 95.7%. The few remaining errors are due to transcription or filtering errors. Extracting pronounced names generates more errors. The main difficulty lies in the transcription and the detection of unknown names (we do not have any a priori knowledge of names that could be pronounced).

Modalities	Precision	Recall	F1-measure
WN	95,7%	97,7%	96,7%
PN	73,5%	50%	59,5%

Table 1: Quality of names extraction, WN: written names, PN: pronounced names, UEM segments only (24 hours of video)

Despite the lower precision and recall of the PN relative to WN, they provide more hypothesis names (see Table 2). We can observe that there are about twice more pronounced names compared to written names, whether we analyze raw videos or UEM only. This proportion is valid for the number of names occurrences or the number of different persons.

Modalities	Segment	#Occurrences of names	#Persons w/o duplicates
WN	UEM (24h)	1407	458
	Raw (58h)	2090	629
PN	UEM (24h)	2905	736
	Raw (58h)	4922	1156

Table 2: Number of written (WN) and pronounced names (PN)

5. Naming Speaker using Pronounced Names (PN) and/or Written Names (WN)

5.1. Metrics used

The number of nameable speakers has been evaluated for each video (mono-video):

$$Np_{mono} = \frac{\text{\#videos where } p \in Phr}{\text{\#videos where } p \in Pr}$$

With :

p : a person
Pr : a set of persons p speaking
Phr : Pr with their names writ./pron.

We also evaluated in a cross-video propagation mode :

$$Np_{cross} = \begin{array}{l} 1 \text{ If } p \in Phr \\ 0 \text{ else} \end{array}$$

In other words, the Np_{cross} of person p is equal to 1 if at least in one video the name of p is written/pronounced when the corresponding person speaks in this video, 0 otherwise.

Overall, for all persons, mono- and cross-video scores are:

$$N_{mono} = \frac{\sum_{p \in Pr} Np_{mono}}{\#p \in Pr} \qquad N_{cross} = \frac{\sum_{p \in Pr} Np_{cross}}{\#p \in Pr}$$

If we look at this example with three videos (V_A, V_B, V_C) and five speakers (S_1 to S_5). The name of these speakers can be written or pronounced in each video (N_1 to N_5) :

	V_A	V_B	V_C
$Speaker:$	S_1, S_2, S_3	S_1, S_3, S_4	S_1, S_5
$Names:$	N_1, N_2	N_3, N_5	N_5, N_4

We obtain these scores for each speaker and for all the speakers:

	S_1	S_2	S_3	S_4	S_5	$Global$
Np_{mono}	1/3	1/1	1/2	0/1	1/1	$N_{mono} = 0.57$
Np_{cross}	1	1	1	0	1	$N_{cross} = 0.8$

S_1 speaks in the three videos but can be named only in video V_1. Therefore the corresponding Np_{mono} is equal to 1/3, and Np_{cross} is equal to 1. The name of S_4 has never been pronounced in the video where he speaks, so this speaker is considered as not nameable (Np_{mono}=Np_{cross}=0).

In addition we also count the **occurrence** number:

Occ : occurrences number of names written/pronounced
Occ_{pv} : #Occ when the corresponding person speaks in the UEM segments

A larger number of occurrences may help the name-person mapping. However, since the manual annotation of the written text is not complete (only one image every ten seconds is annotated in average), Occ_{pv} is probably under-evaluated but can be used at least to compare the potential of WN versus PN.

In the following tables, we use the following notations:

M_{uem} : manual annotations on UEM segments
A_{uem} : automatic annotations on UEM segments
A_{raw} : automatic annotations on raw videos
WN : written names
PN : pronounced names

5.2. Unsupervised Speaker Naming

In Table 3, we observe that the written names extracted automatically can name 73.5% of the 555 speakers. The manual annotation of WN is not complete (1 image / 10 sec only), which explains the higher score of the automatic system (73.5%) compared to manual annotations (60.5%). The combined use of the two modalities (WN+PN) enhances the score (+19.9 % in the case of manual annotations - M_{UEM} but fewer when automated systems are used (+2.3 % for A_{uem})). A cross-video propagation increases the N_{cross} approximatively by 4% on average.

The use of the raw videos (A_{raw}) increases the occurrences number of speakers name (Occ_{pv} from A_{uem} = 2262 to A_{raw} = 2781) without significantly increasing the number of speakers nameable in the UEM segments (A_{uem} = 75.8 % to A_{raw} =

76.9 %). But, the additional occurrences of names may facilitate the name-person mapping.

Finally, it is important to mention that the percentages of Occ_{pv} for A_{raw} are underevaluated, ground truth annotation involving only UEM segments. So we can not say if names do match to a person speaking outside the UEM segments .

PN	WN	Occ	Occ_{pv}	N_{mono}	N_{cross}
M_{uem}	-	4273	1863 (43,6%)	62,2	66,5
-	M_{uem}	1049	1022 (97,4%)	60,5	65,9
M_{uem}	M_{uem}	5322	2885 (54,2%)	80,4	83,6
A_{uem}	-	2905	914 (31,5%)	26,7	30,8
-	A_{uem}	1407	1348 (95,8%)	73,5	76,8
A_{uem}	A_{uem}	4312	2262 (52,5%)	75,8	78,7
A_{raw}	-	4922	1104 (22.4%)	27.9	32.3
-	A_{raw}	2090	1677 (80.2%)	74.8	77.5
A_{raw}	A_{raw}	7012	2781 (39.7%)	76.9	79.3

Table 3: Mono- and Cross-video scores, and number of occurrences of named speakers, for PN and WN modalities - 555 manually annotated spkrs - 24h (UEM) or 58h (Raw) of video

5.3. Detail per speaker's role

In the REPERE corpus, five different speaker categories have been defined to classify people (anchor, columnist, reporter, guest, other). In view of the detailed results, we merged categories with similar behavior for a better readability. The first three were grouped into the role R1: anchor/journalist, the last two in the role R2: guest/other. Table 4 shows the speaker distribution according to their roles. A role has been assigned to each person identified in the videos, a person may have different roles depending on the show. Speakers of R1 cover 45% of speech time while they represent only 15% of the speakers.

Role	#Speakers	Speech Time	#Speech Turn
R1	84 (15%)	632 (45%)	6149 (42%)
R2	475 (85%)	783 (55%)	8378 (58%)

Table 4: Speakers distribution according to their roles. **R1**: anchor/journalist, **R2**: guest/other.

Table 5 details the speakers' nameability depending on the role:

PN	WN	Occ_{pv}		N_{mono}		N_{cross}	
		R1	R2	R1	R2	R1	R2
M_{uem}	-	414	1449	79.4	59.0	86.9	62.7
-	M_{uem}	91	931	23.5	66.6	35.7	70.7
M_{uem}	M_{uem}	505	2380	81.5	80.0	89.3	82.3
A_{uem}	-	58	856	13.9	28.7	16.7	33.1
-	A_{uem}	174	1174	38.3	79.3	47.6	81.5
A_{uem}	A_{uem}	232	2030	43.3	81.1	52.4	82.9

Table 5: Mono- and Cross-video scores, as well as number of Occurrences of named speakers, for both PN and WN according to roles (**R1**: 84 anchor/journalist, **R2**: 475 guest/other)

We can see that the name of the 84 anchors/reporters are relatively rarely pronounced (Occ_{pv} to M_{uem} = 414, A_{uem} = 58) or written (Occ_{pv} to M_{uem} = 91, A_{uem} = 174). Indeed, anchors/journalists are often cited by their first names. In addition, their names are difficult to transcribe because they may be unknown to the automated systems (we do not use a priori knowledge of the anchor/reporters names). People in R1 are quite difficult to automatically name while they represent 45% of the speech time. Concerning the percentage of nameable speakers, 79.4% of people in R1 have their names pronounced but only 13.9 % can be retrieved automatically. People in R2 seem to be more automatically nameable than those of R1 (WN +41%, PN +14.8%). The combined use of both modalities (PN+WN) can increase the nameable speakers and occurrences number whatever the type of role taken into account or the propagation mode (mono- or cross-videos).

5.4. Name-to-Speaker Mapping: Oracle Experiments

Until now, we collected percentage of nameable speakers and number of occurrences figures which gave us an idea of the potential of both WN and PN modalities for unsupervised naming of persons in video. But the name-to-speaker mapping step was not considered yet. State of the art systems which do that are currently restricted to adjacent speech turns to affiliate a name to a speaker. In this section, we compare the mapping (affiliation) ability of names written or pronounced to the correct speaker with the help of an oracle. For written names (WN), the oracle considers that the name-person mapping is correct if he speaks during the display of the name. For pronounced names (PN), the oracle considers that the mapping is correct if the right person speaks during the current, previous or the next speech turn.

PN	WN	Oracle(% spk correctly named)	
		mono	cross
M_{uem}	-	53.4	58.9
-	M_{uem}	60.2	65.6
M_{uem}	M_{uem}	76.9	80.5
A_{uem}	-	21.3	25.9
-	A_{uem}	72.9	76.0
A_{uem}	A_{uem}	75.2	78.4

Table 6: Oracle Name-to-Speaker Mapping Performance for the 555 speakers (UEM segments only - 24h video).

The results in table 6 are to be compared with those in table 3. When the affiliation is restricted to adjacent speech turns, the performance reduces (in absolute) from 4.9 % to 8.8 % depending on the system used and on the propagation considered. The reduction is, however, less important in the case of written names (reduction from 0.3 % to 0.8 %). Despite this, the table shows the complementarity of both PN and WN modalities.

6. Conclusion

The pronounced names (PN) and written names (WN) on the screen are an important source of information to name people in broadcast TV. Despite a larger number of PN in the manual speech transcription in our video corpus, speech transcription and named-entities errors reduce the potential of this modality for naming speakers. On the contrary, with our WN detection and transcription system, we were able to obtain twice nameable speakers as the one obtained using PN. Also, it is worth mentioning that the mapping (affiliation) of WN to the right speakers is inherently simpler than for PN. Despite these differences, both methods were shown to be complementary and unsupervised multi-modal name propagation should be developed in the future to improve speaker indexing of TV shows.

7. References

[1] Canseco-Rodriguez L., Lamel L., Gauvain J.-L., Speaker diarization from speech transcripts , the 5th Annual Conference of the International Speech Communication Association, INTERSPEECH, 2004, p. 1272-1275, Jeju Island, Korea.

[2] Canseco L., Lamel L., Gauvain J.-L., A Comparative Study Using Manual and Automatic Transcriptions for Diarization , IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU, 2005, p. 415-419, Cancun, Mexico.

[3] Tranter S. E., Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio , the 31st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2006, p. 1013-1016, Toulouse, France.

[4] Mauclair J., Meignier S., Estève Y., Speaker diarization : about whom the speaker is talking? , IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop, 2006, p. 1-6, San Juan, Porto Rico.

[5] Estève Y., Meignier S., Deléglise P., Mauclair J., Extracting true speaker identities from transcriptions , the 8th Annual Conference of the International Speech Communication Association, INTERSPEECH, 2007, p. 2601-2604, Antwerp, Belgium.

[6] Jousse V., Petit-Renaud S., Meignier S., Estève Y., Jacquin C., Automatic named identification of speakers using diarization and ASR systems , the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2009, p. 4557-4560, Taipei, Taiwan.

[7] Poignant J., Bredin H., Le V.B., Besacier L., Barras C., Quénot G., Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast , the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH, 2012, Portland, USA.

[8] Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., The REPERE Corpus : a Multimodal Corpus for Person Recognition , the 8th International Conference on Language Resources and Evaluation, LREC, 2012, p. 1102-1107, Istanbul, Turkey.

[9] Poignant J., Besacier L., Quénot G., Thollard F., From Text Detection in Videos to Person Identification , IEEE International Conference on Multimedia and Expo, ICME, 2012, p. 854-859, Melbourne, Australia.

[10] Lamel L. et al., Speech Recognition for Machine Translation in Quaero , The International Workshop on Spoken Language Translation, IWSLT, 2011, San Francisco, USA.

[11] Gauvain, J.L., Lamel, L., Adda, G., The LIMSI Broadcast News Transcription System , Speech Communication, 2002, v. 37, p. 89-108.

[12] Gauvain, J.L., Lamel, L., Adda, G., Partitioning and Transcription of Broadcast News Data , the 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, ICSLP, 1998, p. 1335-1338, Sydney, Australia.

[13] Dinarelli M. and Rosset S., Models Cascade for Tree-Structured Named Entity Detection , the 5th International Joint Conference on Natural Language Processing, IJCNLP, 2011, p. 1269-1278, Chiang Mai, Thailand.

[14] Lavergne T., Cappé O., and Yvon F., Practical very large scale CRFs , the 48th Annual Meeting of the Association for Computational Linguistics, ACL, 2010, p. 504–513, Uppsala, Sweden.

[15] Allauzen A. and Bonneau-Maynard H., Training and evaluation of pos taggers on the french multitag corpus , the 6th International Conference on Language Resources and Evaluation, LREC, 2008, Marrakech, Morocco.

[16] Bernard G., Rosset S., Galibert O., Bilinski E., and Adda G., LIMSI participation in the QAst 2009 track , the 10th Workshop of the Cross-Language Evaluation Forum, CLEF, 2009, p. 289-296, Corfou, Greece.