DETECTING PERSON PRESENCE IN TV SHOWS WITH LINGUISTIC AND STRUCTURAL FEATURES

Frederic Bechet, Benoit Favre

Aix Marseille Univ LIF/CNRS
Marseille, France
{frederic.bechet,benoit.favre}@lif.univ-mrs.fr

Geraldine Damnati

France Telecom - Orange Labs
Lannion, France
geraldine.damnati@orange.com

ABSTRACT

Person detection and recognition in videos is a hard problem due to the intrinsic ambiguities of the sound and image channels and their interaction. Whatever method is used to extract person hypotheses from the audio or the image channels, person recognition in videos relies on a multimodal decision process that merges the different hypotheses produced in order to decide, for each frame, who is present in the video at the audio level, at the image level or at the content level (person mention in speech or inserted text boxes). In this framework the focus of this paper is to produce a list of person presence hypotheses from the audio channel of a video document only, to be used in addition to person presence detected at the image level by a multimodal fusion process. In this study we focus on the audio channel only, using two kinds of features: *linguistic features* corresponding to the way a person is mentioned by a speaker; *structural features* corresponding to the context of occurrence of a name in a show. We show that both sets of features are complementary and that good results can be achieved on a TV show corpus annotated with person presence labels.

***Index Terms*—** Identification of persons, Named Entity, Boosting, Spoken Language Understanding

1. INTRODUCTION

Multimedia document indexing is an important step toward the efficient use of the huge collections of audio and video documents available through the internet. Among all the possible index features that can be associated with a video document, person identification features are crucial as they allow advanced search queries such as "*find all the video documents talking about M. X*", "*find all the video documents where M. X is talking*" or "*find all the video documents featuring M. X*". These three examples of queries illustrate the different kinds of *presence* and *mention* of a person *X* in a video: *X* is mentioned (by his name or a description) either orally or in a written form in a text box inserted in the images; *X* is one of the speakers; *X* is visible and recognizable in the video images. Of course these three situations are not mutually exclusive.

Person detection and recognition in videos is a difficult problem due to the intrinsic ambiguities of the sound and image channels and their interaction. The audio channel can contain interactive and simultaneous speech, with possibly background noise, leading to ambiguities in the speaker diarization process. Face recognition in the video channel is a difficult task because of the large range of variations in the images representing a person's face, such as variation in camera quality, lighting, background, clothing, hairstyles, pose, expression, etc. Moreover the association of a voice and a face is not straightforward as an image can contain several faces and even not necessarily the speaker's face [1].

Whatever method is used to extract person hypotheses from the audio or the image channels, person recognition in videos relies on a multimodal decision process [2, 3, 4] that merges the different hypotheses produced in order to decide, for each frame, who is present in the video at the audio level, at the image level or at the content level (person mention in speech or inserted text boxes). In this framework the focus of this paper is to produce a list of person presence hypotheses only from the audio channel of a video document, to be used in addition to person presence detected at the image level by a multimodal fusion process.

The method proposed has been developed for the French *REPERE* (person recognition in video) challenge. It consists in extracting from the speech transcriptions of a TV show a list of people names mentioned by the speakers, then deciding for each name if the corresponding person appears in the audio and/or the images of the video. In this study we use two kinds of features: *linguistic features* corresponding to the way a person is mentioned by a speaker; *structural features* corresponding to the context of occurrence of a name in a show (how many time is this name repeated? by what kind of speaker? with what kind of speech?).

This paper is structured as follows: section 2 presents the previous studies that have been carried on the person recognition task both on the audio and image channels; section 3 presents the video corpus used in the experiments and the annotation performed on the speech transcriptions; finally section 4 describes a first evaluation of a person presence detector using only linguistic and structural features extracted from the audio signal of a video TV show.

2. RELATED WORK

Two kinds of methods can be used to perform person recognition in videos: using face or voice dictionaries in order to perform a *verification* of a person's identity given a new sample (face or voice) or extracting a person's identity directly from the video thanks to speech transcription and text inserted in the images. These are of course complementary approaches, as person identification hypotheses output directly from the video signal can be filtered thanks to dictionaries and dictionaries can be dynamically enriched with new persons detected in a document being processed.

Dictionary based methods are known as Speaker Identification methods when using only the audio channel and Face Recognition on the image channel [5]. Since the method proposed in this paper is not based on dictionaries, this study relates more to the sec-

This work is partially funded by the French ANR project PERCOL

ond kind of methods that directly extracts people identities from the video signal. Like dictionary based methods, some methods use only the audio channel [6, 7]: a speaker diarization process is performed, then the possible speaker identities that can be found in the speech transcript are matched to each speaker detected based on linguistic features such as key-phrases (e.g. "*Let us listen to M. X*") and global constraints (a speaker should always have the same identity).

Methods using the image channel are roughly based on the same principle, introduced by the early work of Satoh and Kanade in 1997 [2] with the *Name-It* system: the association between names and faces in news videos is based on the co-occurrence between the detected faces and the names extracted from the transcript. A first clustering process on the faces detected is performed and the name given to each cluster is the one occurring the most often in all the portions of video containing the faces of the cluster. This method has been improved by adding constraints to the face/name association method derived from *a priori* knowledge on the document to process. For example [3] apply this method to identify individuals in news video monologues; [4] uses more constraints by identifying every character of an episode of a popular TV program. In all these studies the possible name candidates are extracted both from the speech transcripts and the text boxes contained in the images and processed by an OCR module.

The task targeted in this study is on one hand a generalization of the person recognition task presented in these previous studies as we are not only interested in recognizing the main participants who feature in a video (with voice and image), but rather in any presence (voice and/or image) of all the persons mentioned; and on the other hand we can consider our work as a component in a multimodal person recognition system, which would associate each name detected in the speech transcript with a confidence score on the possible presence of the person in the video. This kind of information can be used in association with existing voice or face dictionaries as well as dynamic methods such as [8] that uses the WEB in order to collect new data for confirming or infirming a hypothesis.

Detecting and characterizing person presence in TV shows are the goals of the French *REPERE* challenge that funds this study. At each frame of a video document a *REPERE* system must answer three questions: who is talking? who is recognizable on the picture? who is mentioned (orally or in a text box on the picture)?

3. A VIDEO CORPUS WITH MULTIMEDIA ANNOTATIONS

The corpus used in this study is made of 21 TV shows collected from 6 French TV channels between October 2008 and January 2009, with a variable length from 10 to more than 40 minutes and a number of different speakers ranging from 10 to 80. Overall, the corpus corresponds to 7.7 hours of speech, for a total amount of 83.5k words uttered by 760 different speakers. TVBN shows are from the main French generalist channels and are considered as a whole. Broadcast Conversation (BC) portions (interviews, live reports) are kept in the analysis and evaluation process. Each show has been first segmented in *sections*, a section being defined as a coherent segment with a given topic, usually starting with an anchor speaker turn and followed by one or several reports and/or an interview. Then each section has been manually annotated in terms of speaker turn segmentation, elocution mode (planned vs. spontaneous speech), speaker name, speaker role and word transcription.

For the purpose of this study *person* named entity mentions have been located and further annotated along two dimensions. The first dimension corresponds to person entity subtypes and is composed of 5 categories: politician, reporter, function/job (ex: *Vice President*), first name only, other (everything else). Table 1 reports the number of annotated person entities along these 5 categories.

Entity type	politician	reporter	job	firstname	other
Occurrences	263	275	200	105	436

Table 1. Distribution of person entity categories.

The second dimension which is of most interest for this study reflects the *presence* or *absence* of the person referred to by the entity mention. The notion of presence refers only to the current video section. If a person is mentioned in one section but only appears in another section it is not considered as present.

Presence is considered with the following distinctions:

- presence in the audio only: the person speaks within the section but is not visible in the video
- presence in the video only: the person is visible but doesn't speak during the current section
- presence in audio and video: the person is visible and speaks within the section (not necessarily simultaneously)
- not present: the person name is just mentioned but does not correspond to a person present in the section

On the overall, 1279 occurrences of person entities have been identified and annotated along these two dimensions. A person entity can occur several times in the same section, thus the number of distinct person entities when restricting to one occurrence per section is 1018. With a total of 243 sections, the average number of distinct person entities per section is 4.2 while the average number of person entity occurrences per section is 5.3.

The following figures are expressed relatively to the total amount of 1279 annotated person entity occurrences in the first column and relatively to the 1018 distinct person entities in the second column.

Label	**# occ.**	**# distinct entities**
presence in the audio only	147	145
presence in the video only	399	300
presence in the audio and video	324	214
not present	362	333

Table 2. Repartition of person entity occurrences in terms of presence.

As presented in table 2, 71.7% of person entity occurrences correspond to a present person. The proportion relatively to distinct person entities is 68.5%. 97% of the person entities that are only present in the audio are actually reporters that are cited and introduced by the anchor speaker. They are usually only mentioned once in a given section. Apart from that category, persons that are present in the video or both in the audio and video are mentioned on average 1.4 times in the same section.

In French TVBN shows, reporter names are usually introduced by the anchor speaker, in the turn preceding the beginning of a report. The anchor speaker, usually introduces both the cameraman and the reporter who will comment the images. Hence only the latter is considered as Present while no particular difference can be observed in the way they are introduced. In our corpus, 59% of reporter names occurrences correspond to Present and 41% are cameramen reporters that are not Present in the given chapter.

4. EXPERIMENTS

4.1. Setup

We present in this section a first evaluation of a person presence detection system using only clues obtained on the speech transcription of a video TV show. In particular, we are interested in discovering if we could take advantage of this unimodal view to complement other views when they fail (when, for instance, a face is occluded). Unlike other studies presented in section 2, we want to detect any kind of person presence: audio, video or both. The main question we want to answer is the following: can we guess, only from speech transcription, if a person mentioned by one of the speakers of a TV show appears, either orally or just visually, in the same section of the show? To answer that question we have built an experimental setup on the corpus presented in section 3. While this corpus was annotated with fine-grained classes denoting audio or video presence, the following experiments are carried on as a 2-way classification problem: *presence* vs *absence*. In particular, for each name mentioned in a given section, an instance is created with label "N" (for *non presence*) if this person does not speak and is not visible in this section and label "P" (for *presence*) otherwise.

A range of classifiers could tackle the job of performing presence label prediction. We pick AdaBoost [9] for its versatility and robustness but other choices should be considered in future work. Let $\mathbf{x}$ be the feature vector representing a given instance, $y \in \{P, N\}$ is the class label that has to be predicted from it. We seek $\hat{y} = argmax_y P(y|x)$, the label of highest probability given a feature vector. Adaboost generates such probability through weighted one-level decision trees on feature values.

$$P(y|x) = \left[exp \left(-2m \sum_{i=1}^{m} w_i^{(y)} s_i(x) \right) \right]^{-1}$$

where $s_i(x)$ is a one-level decision tree over a single feature (presence of a word n-gram, threshold on a real-valued feature...), $w_i^{(y)}$ is the associated weight for class y. The classifier is trained by greedily searching for the m best decision-trees over the training set. In the following experiments, we use the icsiboost [10] implementation of Adaboost.

It is difficult for this task to guess in advance what will be relevant features for presence prediction. In TV shows, since both visual and audio modalities are available to the audience, speakers don't necessarily specify that a person previously mentioned is visible or is talking. Therefore, we look in this study at an exploratory set of features that both encompass linguistically motivated features, as well as structural features. Linguistic features include word n-grams around each occurrence of a person name, the name itself, the verb phrase located just after each name occurrence to model phenomena such as "John Doe *reports* about..." and a marker when the name is located at the end of a turn. Structural features consist in the duration of the section, the number of times the name is referred to in the section (including subparts of the name, such as the family name), the number of turns it contains, the role of the speaker and next speaker of each occurrence of the entity (anchor, reporter, other) and whether the speech is spontaneous or planned.

Given this set of features, we performed a leave-one-out experiment at the TV show level by using, at each iteration, one TV show for the test, one for parameter tuning, and the remaining for training. Since we wanted to check in this first study the feasibility of the task, we obtained our features in these experiments from the reference transcriptions and labels presented in section 3 with the list of manually annotated person names.

4.2. Results

We report in these results the error rate, precision, recall and F-score obtained on our corpus for each class "*P*" and "*N*". We kept from the corpus 717 instances of person names (those referring to a specific person) for which a presence label has to be predicted, among which 270 are not present (*N*) and 447 are present (*P*). A baseline system considering that all persons are present would achieve a precision of 62.34% with a recall of 1, and an F-score of 76.80%. This baseline is hard to beat given the distribution of classes *P* and *N* and the intrinsic difficulty of the problem. However our goal is to improve the precision in the detection as this will be an important feature in a multimodal decision process assessing the presence of a person at a certain time in a video document.

System results are reported in Table 3. We see that while it fails to beat the baseline in term of F-score, its precision is higher. This behaviour is favorable when contemplating multimodal decisions because higher precision will lead to less noise fed to the integration component. We also look at the impact of linguistic and structural features. An interesting finding lies in the error-rates of the *N* classes that are much higher than that of the combined system which seems to correct that imbalance.

Features	**Class**	**Err**	**Prec**	**Rec**	**F-score**
All	N	66.30	54.82	33.70	41.74
	P	16.78	67.51	83.22	74.55
Linguistic only	N	88.15	40.00	11.85	18.29
	P	10.74	62.64	89.26	73.62
Structural only	N	77.04	56.88	22.96	32.72
	P	10.51	65.79	89.49	75.83

Table 3. Leave-one-out performance, in percent, of the system on different feature subsets. Error rate (Err), precision (Prec), recall (Rec) and F_1-score (F-score).

Figure 1 shows the tradeoff between precision and recall that can be tailored to a specific application. While the rather recall-oriented default decision boundary favors the combination of all features, structural features might yield better precision if this measure is favored over recall. Linguistic features consistently underperform structural features, showing that word ngrams do not capture person presence indicator very well, probably because not enough training data is available to extract good patterns.

Table 4 lists the spread of model weights among feature categories, computed for an average test example. The weight for category c is given by:

$$\mathcal{W}_c^{(y)} = \sum_{i \in c} \sum_{x \in t} w_i^{(y)} s_i(x)$$

where t is the test set and y is label P. This table shows that speaker role is the most prominent feature, which is expected as people are mentioned differently according to the role of the speaker. Then, word n-grams have the second highest weight, expectedly, followed by other lexical features (name text, and action verb), and finally other structural features. When looking at feature subsets, it is interesting that the number of occurrences of a name yields a much higher weight than when used together with linguistic features.

4.3. Discussion

The main difficulty for detecting automatically the presence of persons in TV shows is that excepting for the anchor and reporters who

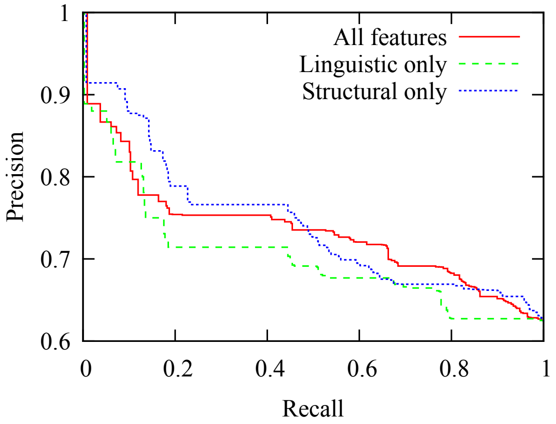

Fig. 1. Precision-Recall curve created by tuning the decision threshold of the classifier.

Feature	All	Ling.	Struct.
Speaker Role	59.16	-	82.90
Word context	24.50	94.70	-
Person name	7.39	3.61	-
Verb	2.67	1.67	-
Sec. duration	4.56	-	0.15
Spontaneity	3.60	-	0.94
Num. occ.	1.37	-	15.99

Table 4. Repartition of model weight for each group of features (in percent of total absolute weight).

are necessarily present, most people mentioned in the speech transcription are only present because of editorial choices. In order to assess this difficulty, we asked two human judges to choose for all name mentions in the corpus whether they were present, absent or whether it was not possible to determine their presence using only the spoken transcription of the turns containing people name. We compared the result of this transcript-only annotation to the gold-standard reference created from the TV recordings. Results are summarized in Table 5 where human choices are compared to audio or audio-video presence and video only presence (as percentages of the total number of decisions). They show that in most cases it is difficult to determine that someone is absent because the transcript would have to explicitly state that the person is not on screen. Then, It is interesting to note that humans have more difficulties identifying video-only presence and do the best job at determining the presence of actual speakers.

Audio-only	Reference		
judgement	Absent	Audio or AV	Video only
Absent	1.60	0.09	0.19
Present	1.42	36.04	9.43
Unknown	24.15	7.26	19.81

Table 5. Repartition of human judgements using transcripts only.

5. CONCLUSION

We have presented in this study a first evaluation of a person presence detection system using only clues obtained on the speech transcription of a video TV show. We wanted to answer to the following question: can we guess, only from speech transcription, if a person mentioned by one of the speakers of a TV show is going to appear, either orally or just visually, in the same section of the show? We have shown that a classifier using linguistic and structural features can achieve good performance on this task, making use of a wide range of features from the speaker role to the verb phrase following the person name occurrence. We need now to validate these results on a fully automated system, using ASR transcription and annotation instead of reference ones in the classifier features, and also to integrate the result of this system to a multimodal decision process that is going to be evaluated during the French REPERE challenge.

6. REFERENCES

[1] M. Bendris, D. Charlet, and G. Chollet, "People indexing in tv-content using lip-activity and unsupervised audio-visual identity verification," in *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*. IEEE, 2011, pp. 139–144.

[2] Shin'ichi Satoh and Takeo Kanade, "Name-it: Association of face and name in video," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 368, 1997.

[3] J. Yang and A.G. Hauptmann, "Naming every individual in news video monologues," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 580–587.

[4] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy–automatic naming of characters in tv video," in *Proceedings of the British Machine Vision Conference*, 2006, vol. 2.

[5] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *University of Massachusetts, Amherst, Technical Report 07*, vol. 49, pp. 1, 2007.

[6] SE Trantee, "Who really spoke when? finding speaker turns and identities in broadcast news audio," in *ICASSP 2006*. IEEE, 2006, vol. 1, pp. I–I.

[7] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin, "Automatic named identification of speakers using diarization and asr systems," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4557–4560.

[8] C. Liu, S. Jiang, and Q. Huang, "Naming faces in broadcast news video by image google," in *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 717–720.

[9] R.E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, vol. 37, no. 3, pp. 297–336, 1999.

[10] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," http://code.google.come/p/icsiboost.