Improving Speaker Identification in TV-shows using person name detection in overlaid text and speech

Delphine Charlet¹, Corinne Fredouille², Géraldine Damnati¹, Gregory Senay²

¹ Orange Labs, Lannion, France
² University of Avignon, CERI/LIA, France
delphine.charlet@orange.com, corinne.fredouille@univ-avignon.fr

Abstract

This paper is dedicated to the use of auxiliary information in order to help a classical acoustic-based speaker identification system in the specific context of TV shows. The underlying assumption is that auxiliary information could help (1) to re-rank n-best speaker hypotheses provided by the acoustic-based only speaker identification system, (2) to provide confidence score to refine a rejection process (open-set identification task), and finally, (3) to identify speakers not covered by the speaker dictionary (out-of-dictionary speakers) used by the speaker identification system (full-set verification task); the last point being one of the main issue when dealing with TV shows. In this paper, the auxiliary information is based on person names detected in overlaid text and speech. Experiments conducted in three different datasets issued from the REPERE evaluation campaign have highlighted the interest of the auxiliary information used here, and notably the use of overlaid person names to identify out-of-dictionary speakers, confirming the key assumptions made.

Index Terms: Speaker identification, overlaid text, broadcast shows, spoken name detection.

1. Introduction

The aim of this study is to identify speakers in TV-shows, thanks to acoustical speaker recognition. As it is impossible to build a voice dictionary of all the people who can appear in TV-shows (with new people appearing every day), the task is actually an open-set speaker identification problem: is the speaker one of the known speakers in the dictionary, if yes, which one, or none of them? Thus it raises the need of a rejection process to decide whether the speaker is one of the speakers in the dictionary or not. Although open-set identification has long been studied, very few studies have focused on broadcast contents. For instance, [1][2][3] study open-set speaker identification on telephone or laboratory databases, controlling a lot of factors: size of the training set, duration of testing utterances, number of speakers in the dictionary. When it comes to identify speakers in broadcast contents, a number of new difficulties arises: first, the signal is not segmented into speaker turns, the amount of speech contents, a number of new difficulties arises: first, the signal is not segmented into speaker turns, the amount of speech duration of testing utterances, number of speakers in the databases, controlling a lot of factors: size of the training set, open-set speaker identification on telephone or laboratory contents, and focused on the discriminative training of the speaker models.

Recently, open-set speaker identification in broadcast-contents has received attention, focusing on other information that can be available to identify speakers in broadcast contents: [5] studied the combination of transcription-based and acoustic-based speaker identification on radio contents, and [6] explores the speaker identification using only overlaid names in TV-contents, without any acoustical model. In our work, we propose to explore the usefulness of speech transcripts and overlaid names to improve acoustic-based speaker identification in TV-contents.

2. Speaker Identification

2.1. Speaker Segmentation and Clustering

The diarization system used in this work is the one presented in [7]: it is a sequential processing using firstly Bayesian Information Criterion and then Cross-likelihood Criterion, with special attention paid for overlapped speech, which is non-negligible in TV-debates. Overlapped speech segments are first detected and discarded from the clustering process, and then reassigned to the 2 nearest speakers, in terms of temporal distance between speech segments.

2.2. Speaker Modeling

Each speaker, belonging to the speaker dictionary associated with the acoustical based identification system, is modeled following a trivial UBM/GMM scheme. Speaker models are derived from the mean adaptation of a 512 Gaussian UBM. Given a TV-show, this speaker dictionary and models associated with are then used to provide acoustic scores based on the output of the speaker diarization system previously described. For a given speaker cluster, likelihood scores are computed between all frames related to the cluster and each model associated with the speaker dictionary and averaged to provide and acoustic score per cluster and model. The system provides the N-best speaker hypothesis per cluster. These scores and speaker hypotheses are attributed to all the segments of the speaker cluster.

2.3. Speaker Scoring

The speaker scoring must fulfill 2 needs: 1) for ranking n-best speaker hypotheses, 2) for rejecting out-of-dictionary speakers. Let X be the speech utterance, λi the hypothesized speaker, and \(s(\lambda_i/X)\) the acoustic score of the hypothesized speaker for the given speech utterance:

\[
\text{Step1: } \lambda^{\text{X}} = \underset{i=1,\ldots,N}{\arg\max} s(\lambda_i/X)
\]

\[
\text{Step2: if } s(\lambda^{\text{X}}/X) \geq \Theta, \text{ accept, else reject}
\]

Hence, various functions can be applied on the acoustical score. Z-norm is a model-dependent transformation, thus it can modify the ranking of the n-best hypotheses. As far as T-norm is concerned, as it is a test-dependent transformation, it applies the same transformation on the segment whatever the speaker hypothesis. Thus, it does not change the ranking but may have an influence on the rejection process. The difference of scores...
between hypotheses of different ranks can also be used. We have computed the following transform, named deltascore in the rest of the paper: for first rank hypothesis, the deltascore is the difference between the score of the first hypothesis and the second hypothesis. For hypothesis of further ranks, the deltascore is the difference of scores between the current hypothesis and the first hypothesis. This deltascore is thus monotonously decreasing according to the rank of the hypothesis and does not change the ranking but may have an interest for the rejection step. We can also apply the deltascore on the T-norm scores.

3. Using Auxiliary Information

3.1. Principle

Apart from acoustical recognition scores, there are several other sources of information that can contribute to identify speakers in TV-show. For instance, names in overlaid text or information extracted from the speech can be used.

We use a general framework to include auxiliary information, in order to compute a new recognition score, which can enable to re-rank the n-best acoustical hypothesis (step 1), and can also be used as confidence measure for the rejection of out-of-dictionary speakers (step 2).

The framework is based on logistic regression:

\[
\text{let } \left( d_{i}^k(X) \right)_{k=1,\ldots,K} \text{ be the set of descriptors associated with the speaker hypothesis } \lambda_i \text{ for speech turn } X. \text{ The fusion score is: }
\]

\[
SF(\lambda_i / X) = \frac{1}{1 + \exp\left(-\left( a_x + \sum_{k=1}^{K} a_k d_{i}^k(X) \right) \right)}
\]

3.2. Overlaid Person Names (OPN)

A multi-stage system has been specifically designed in order to detect and recognize Overlaid Person Names (opn) [7]. Text Detection is achieved on each frame and each text order to detect and recognize Overlaid Person Names (opn) from which the most probable sequence is extracted. A rule-based classifier (based on the length, position, number of words in the box ...) decides whether the text is an Overlaid Person Name and GnuOCR) on each frame of the track, the resulting representation of the candidate names and the phonetic confusion network of the ASR.

When a speaker is out-of-dictionary in the acoustic, which processes in the following way:

4. Identifying out-of-dictionary speakers

When a speaker is out-of-dictionary in the acoustic modeling, it is still possible to identify him/her thanks to auxiliary information such as opn and spoken name. We propose a hierarchical combination rule, using only opn and acoustic, which processes in the following way:

Preliminary step:

Opn candidate for each speech turn X: opn name(X) is the opn which has maximal overlap duration with X, may not exist (no opn during the speech turn).

Opn candidate for each cluster: opn name(Cluster) is the opn which has maximal overlap duration with the cluster, may not exist (no opn during the cluster).

Naming each speech turn X:

Foreach X:

\[
\lambda_x = \text{argmax}(SF(\lambda / X))
\]

If \( SF(\lambda_x / X) > \Theta \) then name(X) = \( \lambda_x \)

Else if opn name(X) exists

then name(X) = opn name(X)

Else if opn name(Cluster) exists

then name(X) = opn name(Cluster(X))
5. Experiments

5.1. Experiments setup

5.1.1. Corpus

The corpus used in this work is provided by the REPERE evaluation campaign [12] whose goal is to develop methods of multi-modal people identification in TV contents. The corpus contains videos (news, political debate and talk-shows) from 7 French TV programs (2 programs from BFMTV channel and 5 from LCP channel) broadcasted between 2011 and 2012. The campaign provides a training corpus (for phase1 in REPERE challenge) Corp1 of 135 videos with total duration 48 hours, from which excerpts totaling 24 hours were annotated. The annotations were performed every 10 seconds approximately, which results in 8766 annotated keyframes. This Corp1 corpus is used as a development set, to tune the systems. Whenever a training pass is necessary, for learning the weights of the logistic regression or training acoustic model for instance, it is performed in a leave-one-out fashion so as to get recognition results on Corp1 set.

2 corpora are used as test sets: Corp2 and Corp3 (test corpus for phase0 and phase1 respectively in REPERE challenge) which consist of 25 and 28 shows resp., resulting in 1108 and 1250 annotated keyframes resp. Moreover, Corp1 corpus (under leave-one-out fashion) is coupled with some additional data issued from French Broadcast news TV shows to estimate the GMM models associated with the speaker dictionary. This dictionary is focused on 345 politics, journalists and anchor people for whom a minimum 30s duration data are available. Thus, data duration used for speaker model estimate can largely vary from one speaker to another, from 30s (minimum value required) to about 3h (some anchor people) with an average value of 253.42s. The length of segments available in these data varies also largely from less than 1s to about 4mn, with an average of 7.7s.

5.1.2. Evaluation Metrics

Due to the particular annotations of the corpus, in terms of keyframes, evaluation is performed in terms of number of occurrences of speakers in the annotated keyframes. Given the fact that some keyframes may contain multiple speakers whereas others may not contain any speaker, the number of speaker occurrences is not equal to the number of keyframes. Table 1 shows the number of speakers and speaker occurrences in the 3 corpora, with the amount of in-dictionary speaker and speaker occurrences and the percentage they represent relatively to the whole set of speaker occurrences. It can be pointed out, especially regarding Corp1 corpus, that the number of keyframes per speaker is varying a lot from one speaker to another as the rate of keyframes covered by the speaker dictionary is largely higher than the one of speakers.

Table 2. Performance (%) of acoustic speaker identification on Corp1

<table>
<thead>
<tr>
<th>Decision score</th>
<th>Pr</th>
<th>Re</th>
<th>Fm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech-turn score</td>
<td>69.0</td>
<td>30.3</td>
<td>42.1</td>
</tr>
<tr>
<td>(Cluster) score</td>
<td>80.1</td>
<td>63.8</td>
<td>71.0</td>
</tr>
<tr>
<td>Z-norm score</td>
<td>79.2</td>
<td>62.1</td>
<td>69.6</td>
</tr>
<tr>
<td>T-norm score</td>
<td>89.6</td>
<td>69.7</td>
<td>78.4</td>
</tr>
<tr>
<td>Deltascore</td>
<td>88.0</td>
<td>70.7</td>
<td>78.4</td>
</tr>
<tr>
<td>deltascore(T-norm)</td>
<td>84.5</td>
<td>72.3</td>
<td>77.9</td>
</tr>
</tbody>
</table>

Performance of speech-turn level score is reported for contrast with cluster-level score (which is the default score in this work): indeed, the recall is very low as many segments are too short (average length 7.7s for speech turn and 86s for cluster) to take a reliable decision. Z-norm doesn’t help, contrary to T-norm and deltascore which enables each an absolute 7.4% F-measure improvement. Unfortunately, the combination of both transforms does not help. In the rest of the work, we use only the deltascore.

Table 3 presents the results obtained when integrating auxiliary information in the fusion score, for Corp1, Corp2 and Corp3. For Corp1, results with Oracle speaker diarization are also presented in Corp1_OracleDia, so as to evaluate the potential of the proposed approach, independently of the performance of the speaker diarization. The Diarization Error Rate of the system used in our experiments is, on Corp1, 8.9% when excluding overlapped speech and 12.1% when including overlapped speech in the evaluations.
First, a general comment on the results across the corpora: results are consistent between Corp1 and Corp3, where consistent improvements due to opn or spoken names are observed, whereas these features do not lead to improvement for Corp2, which already achieves very good results merely on acoustical speaker recognition. Hence, these very good results on Corp2 are only slightly improved using other information. The results presented in the table for spoken names are the best obtained with spoken names based descriptors: using log(1+spoken_aroundclu(λ)), log(1+spoken_insideclu(λ)), which improves the F-measure from 78.4% to 79.6% on Corp1. An interesting point concerns the weights of the predictors: the weights trained on Corp1 corpus are +2.19 for spoken aroundclu and -1.72 for spoken insideclu. As expected, pronouncing a name is a negative cue for being this person.

Table 3. Performance of speaker identification using auxiliary information: F-measure (Prec/Rec)

<table>
<thead>
<tr>
<th>Decision score</th>
<th>Corp1</th>
<th>Corp1 OracleDia</th>
<th>Corp2</th>
<th>Corp3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deltascore</td>
<td>78.4 (88.0/70.7)</td>
<td>83.7 (94.4/75.1)</td>
<td>84.3 (91.0/78.5)</td>
<td>75.5 (84.0/68.6)</td>
</tr>
<tr>
<td>Deltascore +spoken</td>
<td>79.6 (89.2/71.9)</td>
<td>85.3 (92.3/79.3)</td>
<td>84.0 (91.9/77.3)</td>
<td>76.2 (84.9/69.0)</td>
</tr>
<tr>
<td>Deltascore +opn</td>
<td>82.3 (91.7/74.8)</td>
<td>89.9 (96.0/82.7)</td>
<td>84.5 (91.8/78.3)</td>
<td>84.0 (87.8/80.5)</td>
</tr>
<tr>
<td>Deltascore +opn +spoken</td>
<td>82.3 (91.6/74.7)</td>
<td>88.9 (96.5/82.4)</td>
<td>85.5 (91.9/79.9)</td>
<td>83.9 (88.3/80.0)</td>
</tr>
</tbody>
</table>

The major improvement is due to the inclusion of opn information, which enables up to almost 9% absolute improvement in F-measure on Corp3. Unfortunately, opn and spoken information don’t combine well in the proposed fusion framework, as the combination of both does not improve performance obtained using opn descriptor only.

In order to differentiate the interest of opn or spoken name in the re-ranking of speaker hypotheses from their interest in the decision score for rejection, contrastive experiments have been done on Corp1 using opn or spoken names only on the first acoustic-based speaker hypothesis. Spoken names seem to be only efficient in the decision threshold, as performance is the same with or without doing the re-ranking (F-measure=79.6%). On the other hand, opn information is very useful for the decision threshold only (F-measure=80.9%), but also for re-ranking speaker hypothesis, as an absolute further gain of 1.4% of F-measure is observed when doing the re-ranking. Looking into details, 202 speaker hypotheses have been modified with the inclusion of the opn information: indeed, 202 speaker hypotheses were done which were not first ranked by acoustical score. Among these 202 modified first speaker hypotheses, 177 are correct, whereas before including opn, only 5 were correct. Finally, we can see in table 3 that all the improvements observed on Corp1 are even reinforced on Corp1 OracleDia. The auxiliary information from opn and spoken names are very dependent from the overlap of these detections with speaker cluster, thus all the precision we can get in these overlaps leads to improve the speaker identification. It encourages to investigate more in the speaker clustering step.

5.3. Full-set Speaker Identification Results

In this case, we are evaluating the speaker identification system proposed in section 4. In this task which is the primary task of the REPERE challenge [12], we are supposed to identify all the speakers present in the corpus, whatever the information the systems use to name speakers. For the hierarchical combination, we fix the threshold value Θ giving maximal F-measure of deltascore+opn on Corp1 set.

For contrast, results using acoustical score (including or not opn) are given: it is the same system as evaluated in table 3, but considering all the speakers and not only those in dictionary: precision is the same, but recall is lowered by the coverage factor. Results using opn only are also given. Results in the table show the interest of the approach of hierarchical combination, relying first on high precision supervised speaker identification (deltascore+opn), with a back-up to unsupervised speaker identification (opn). In [6], the combination was done just the other way round: first opn, and then acoustic-based speaker identification for unnamed speaker by opn. When we implement this method, we obtain a maximal F-measure of 74.8% on Corp2 (which is the corpus on which [6] report results), instead of 81.1% using our proposed combination scheme.

The hierarchical combination we have proposed succeeds in improving a lot the recall rate, with a complementary coverage of both systems (opn or acoustic system), keeping almost the high level of precision of the acoustic system.

Table 4. Full-set speaker identification: F-measure (Prec/Rec)

<table>
<thead>
<tr>
<th>Decision score</th>
<th>Corp1</th>
<th>Corp1 OracleDia</th>
<th>Corp2</th>
<th>Corp3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deltascore</td>
<td>64.8 (88.0/51.3)</td>
<td>70.1 (91.0/57.0)</td>
<td>61.1 (84.0/47.9)</td>
<td></td>
</tr>
<tr>
<td>Deltascore +opn</td>
<td>68.0 (91.3/54.2)</td>
<td>70.2 (91.8/56.8)</td>
<td>68.6 (87.8/56.3)</td>
<td></td>
</tr>
<tr>
<td>Opn</td>
<td>73.4 (82.0/66.5)</td>
<td>65.5 (78.4/56.2)</td>
<td>70.1 (82.8/60.8)</td>
<td></td>
</tr>
<tr>
<td>Hierar. Combination</td>
<td>83.4 (87.8/79.5)</td>
<td>81.1 (89.3/74.1)</td>
<td>81.9 (87.6/77.0)</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion

This paper has addressed the interest of person name detection in speech and overlaid text to improve a classical acoustic-based speaker identification in the context of TV-shows. Descriptors derived from these detections are combined with acoustical score to re-rank the N-best speaker hypotheses and to provide a score to reject out-of-dictionary speakers. In addition, a hierarchical combination based on both acoustic information and overlaid person names has been designed to identify out-of-dictionary speakers. Experiments on the corpora provided by the REPERE challenge, composed of about 190 TV-shows, have shown significant improvement, especially due to overlaid person names. The gain due to person name detection in speech was less significant, but, considering that speech carries important information about speakers, further investigation will focus on more complex spoken content descriptors, like, for instance, the role of the person who utters the name or the linguistic context of apparition of the name.

This work was partially funded by the French Research Program ANR Project PERCOL in REPERE Challenge.
7. References


