Speakers identification in broadcast TV: facilities and barriers

Author Name[1], *Co-author Name*[2]

[1]Author Affiliation
[2]Co-author Affiliation
author@university.edu, coauthor@company.com

Abstract

...
Index Terms: speaker recognition, error analysis

1. Introduction

2. Notation

Notation	Description
$SpkShow$	All speech segments of a speaker in a video
$SpkSeg$	A speaker turn
T_i^{ref}	the total duration of the speech of $SpkShow_i$, in the reference
T_i^{test}	the total duration where $SpkShow_i$ is recognized by the automatic system
T_i^{corr}	the total duration of correct identification of $SpkShow_i$

3. Systems description

brief description of the systems: training data, modelling, decision...

3.1. PERCOL

3.2. QCOMPERE

3.3. SODA

4. Performance analysis

We are interested here in analyzing the performances obtained per speaker, according to their characteristics, for instance in terms of speech turns etc. As the speech turns properties depend on the show in which the speaker appears, one speaker in one show is considered as the unit of analysis, the so-called $SpkShow$. One speaker appearing in 2 different videos is considered as 2 distinct $SpkShow$.

The test corpus contains about 10 hours of annotated contents, on 62 different videos, totalizing 477 non-anonymous different $SpkShow$.

In the analysis, we adopt the point of view of the references: for each $SpkShow_i$ in the reference is computed a performance measure of the biometric system, defined as the F-measure of the detection of $SpkShow_i$. More precisely, considering the definition given in 2, Precision and Recall can be computed for each $SpkShow_i$:

- $Precision_i = \frac{T_i^{corr}}{T_i^{test}}$
- $Recall_i = \frac{T_i^{corr}}{T_i^{ref}}$
- $Fm_i = \frac{2*Precision_i*Recall_i}{Precision_i+Recall_i}$

Thus, $Fm_i = 0$ means that $SpkShow_i$ was never correctly identified, whereas $Fm_i = 1$ means that $SpkShow_i$ is perfectly identified, without miss detection nor false alarm.

The table 1 shows the average Fm per $SpkShow$ for the different systems.

From the table we can notice the important number of $SpkShow$ which are not in the dictionary of the system, about 40% for each system. As they don't have any model, they obviously cannot be identified, leading to an average global Fm rather poor. More interestingly, the number of $SpkShow$ which are actually in the dictionary and which are not recognised at all, is not negligeable: their represent between 23.5% to 31.5% of the in-dictionary $SpkShow$, according to the systems.

The figure 1 plot the distribution of all the $SpkShow$ in the system dictionaries, according their performance Fm, for the different systems. Foreach $Spkshow$, the average performance and the maximal performance obtained across systems are computed, and the corresponding distribution are also plotted. We can see from this figure that the average performance (from 61.9% to 68.9% according to the systems) presented in table1 is not at all representative of the performances obtained foreach $SpkShow$: speakers are either not recognized or well recognized. Indeed, if we compute the average performance for $SpkShow$ which have $Fm \neq 0$, the average Fm grows to 87.9% for PERCOL, 89.5% for QCompere and 90.3% for SODA.

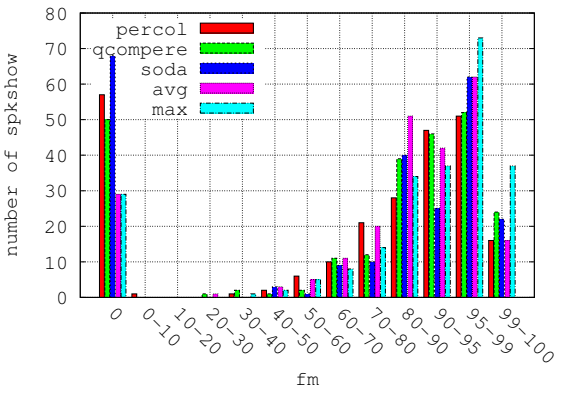

Figure 1: $spkShow$ performance distribution, for each system

To evaluate the impact of the automatic speaker diarization, we also perform the speaker analysis performance, for systems applied on reference speaker diarization. Results for systems PERCOL and SODA are plotted in figures 2 and 3.

	Percol	Qcompere	Soda
average Fm	0.361	0.381	0.351
average Fm for in dictionary speakers	0.628	0.684	0.619
#$SpkShow$ out of dictionary	200	209	204
#$SpkShow$ in dictionary	277	268	273
#$SpkShow$ in dictionary, with $Fm = 0$	79	63	86

Table 1: Average system performances per $SpkShow$

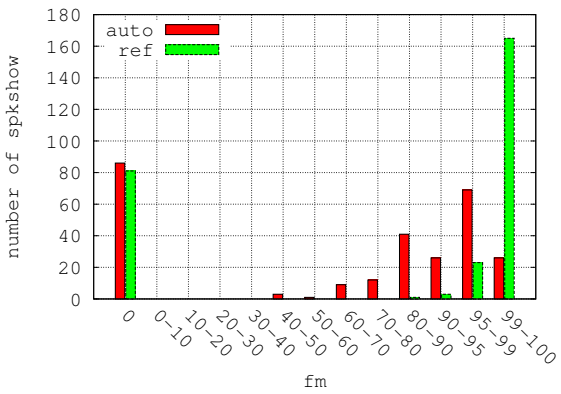

Figure 2: $spkShow$ performance distribution, for SODA system, with reference and automatic speaker diarization

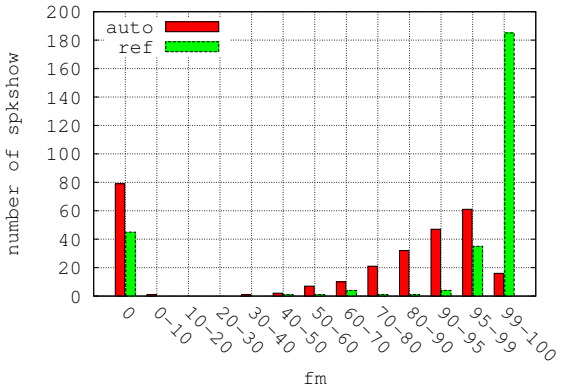

Figure 3: $spkShow$ performance distribution, for PERCOL system, with reference and automatic speaker diarization

5. Performance Prediction

In this section, the aim is to predict the performance of the speaker, from his characteristics in terms of training data and speech turns properties. If we are able to predict, reliably, if the $Spkshow$ will be correctly recognized or not, when analysing on the main features contributing to this prediction, we can identify what are the features that facilitates or hamper the identification, for a given system.

At each $SpkShow$ is associated the maximal Fm obtained accross systems. Doing so, we do not focus on a particular system, but we try to explain "the-best-we-can-do" performance for each $SpkShow$.

5.1. Detection

5.2. Prediction

6. Cross-show extension

7. Conclusions

8. Acknowledgements

9. References

[1] ....