SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
Patricia Besson, Gianluca Monaci, Pierre Vandergheynst, Murat Kunt

ELD 241 (Bâtiment ELD)
Station 11
CH-1015 LAUSANNE

Tel: +41 21 693 5646
Fax: +41 21 693 7600
e-mail: patricia.besson@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Experimental evaluation framework for speaker detection on the CUAVE database

Patricia Besson, Gianluca Monaci, Pierre Vandergheynst, Murat Kunt

École Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-03.2006

January 19, 2006

Experimental evaluation framework for speaker detection on the CUAVE database

Patricia Besson, Gianluca Monaci, Pierre Vandergheynst, Murat Kunt
École Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland
E-mail: {patricia.besson, gianluca.monaci}@epfl.ch

I. Introduction

The goal of this report is to establish an experimental framework to evaluate speaker detection methods on audio-video databases like the CUAVE database [1]. This framework aims at possibly becoming a standard, at least for the institute use, and to be used in future publications.

In a first part, the CUAVE corpus is shortly presented. The second part briefly reviews two speaker detection evaluation frameworks taken from the literature. The main advantages and drawbacks of these experimental methods are exposed before discussing possible solutions in the third part. Finally, an evaluation methodology based on the previous argumentation is proposed.

Notice that the analyzed frameworks, those taken from literature as well as the one proposed here, are dedicated to detection methods requiring a temporal analysis window on which the detection is performed. The duration of this analysis window is considered of t frames or seconds and is shifted by a given value of s frames or seconds.

II. Description of the CUAVE database

The CUAVE speech corpus [1] is a moving-talker speaker-independent database, designed to aid researchers in multimodal speech processing. 36 individual speakers and 22 speaker pairs utter continuous, connected and isolated digits. The sequences present two distinct parts of different complexity. In the first part, the speakers remain still, with only some small, natural motions. In the last part, to the contrary, the speaker moves around intentionally in the individual sequences and both persons are speaking simultaneously in the multiple speaker sequences.

The individual sequences are about 2 minutes long, and the group ones about 20–25 seconds long. The NTSC video standard was used (29.97fps) and the stereo audio signal was sampled at 44kHz.

III. Existing evaluation frameworks

To the best of our knowledge, the CUAVE database has been used to test speaker detection algorithms in two cases [2], [3]. This database has been more widely used in the context of speech recognition, but due to the difference between the two problems, the evaluation frameworks are not directly applicable to the speaker detection case.

The goal of speech recognition is, as the name suggests, that of recognizing the words uttered by a speaker using acoustic and/or visual information. Thus, the evaluation of speech recognition algorithms requires a precise ground truth of the sequences, including the words pronounced, as well as a precise evaluation procedure. On the other hand, the goal of a speaker detection algorithm is that of detecting the person that is currently speaking, in a multi-speaker environment or in adverse conditions. Thus, no information about the content of the speech is needed, and the evaluation framework can be more relaxed than in the previous case.


Fig. 1. Two examples of sequences involving two persons from CUAVE.

Since the framework we are introducing is devoted to the evaluation of speaker detection algorithms performances, we consider, as it was done in [2], [3], only the multi-speaker partition of the database. Each of the 22 clips involves two speakers arranged as in Fig. 1, taking turn in reading series of digits. At the end of the clips, both subjects speak simultaneously reading different sequences of digits. The final part of each sequence has been discarded, and only the parts on which a single person is speaking are considered.

Nock *et al.* present in [2] an empirical study of speaker detection based on audio-visual synchrony. Tests are carried out on the multi-speaker portion of the CUAVE database using the experimental protocol described in the following.

The speaker detection function is based on mutual information and therefore requires a static analysis on a given temporal window. This temporal window is $t = 2$ seconds long (*i.e.* 60 frames) and it is shifted by $s = 1$ second (*i.e.* 30 frames) along the sequence. The results are compared to the ground truth at the center point of each temporal window: the estimates are therefore scored at one second intervals through each clip. Thus, performing their tests on 12 sequences, they end up with a total of 252 test points.

In [3], the authors also used the multi-speaker part of CUAVE to test their speaker detection method. Since their method also requires the use of temporal windows, they define an analysis window of length 60 frames, which is shifted by 20 frames. They eventually end up with 273 test points corresponding to the last frame of the analysis windows.

Clearly, the main advantage of such methods is that the pre-processing time is reduced. The ground truth points where the detection function must be evaluate are in fact easily and quickly established.

However, four main drawbacks can be identified for these two approaches:

- The evaluation function may not be accurate enough. Since the truth about the current speaker for a 1 second interval is given from a unique frame, this make evaluation not very reliable: what happens if speaker 1 is mostly speaking over the current temporal window, but speaker 2 is actually speaking at the observed time instant? Or if nobody is speaking at the considered time instant?
- From the detection algorithm point of view, it does not seem much reasonable to consider 2 seconds of information to perform the detection, and to compare the result to a single 1/60th of the information in the ground truth. Taking again the previous example, a good detection algorithm should indicate speaker 1 as the active speaker if it is the most active on the whole window. But it may happen that the ground truth will indicate such a result as false.
- The evaluation results are very sensitive to the choice of the ground truth and to the choice of the speaker detector's parameters. If, for example, the detection is evaluated not on the central frame of the analysis window but on the next one, the results may be very different.
- Notice that choosing the central frame of the analysis window for the evaluation, somewhere means that the method needs the past and future frames to perform the current speaker detection.

IV. Possible solutions and limits of these solutions

The first point to consider is how to establish the ground truth to which the detector outputs should be compared, in order to assess the detector performances. Let us just recall here that we only consider speaker detection methods where an analysis window is required.

Two ways of establishing the ground truth can be imagined:
1. **Frame level:** each frame is labeled with the current speaker label.
2. **Window level:** each of the t frames that constitute the analysis window are labeled with the label of the person speaking the majority of the time during this period.

The first point to stress is that establishing the ground truth is a tedious and uneasy task, whatever is the chosen approach. The audio and video signals are not perfectly aligned: movements appear in the mouth region before any sound can be heard (co-articulation effect). Thus the starting and stopping frames are difficult to be labeled.

From this point of view, the window level ground truth is much more sensitive to the choices made in the labeling step: let us just consider that a change of speaker occurs in the middle of a period of t frames. If each person is speaking approximately for the same amount of time, it is then a big issue to decide which one is the dominant speaker, and this is strongly dependant on how the labeling has been done. We might consider giving more weight to the speaker already labeled as the dominant one in the previous period.

On the other hand, it must be noticed that the detection algorithm is not performing at the frame level. Then, if choosing to establish the ground truth at the frame level, the evaluation of the method can be performed at a too high resolution.

The possible ways to constitute the ground truth have been discussed. Now we have to consider the evaluation criterion, *i.e.* the way the detector outputs are compared to the ground truth. Once again, different approaches may be considered.

If the window level ground truth is used, the comparison is straightforward: the output of the detector for a given analysis window is compared to the ground truth established for this set of frames. Detection and evaluation are performed on the same temporal window, for the same duration. However, the detector's analysis windows and the ground truth windows should perfectly overlap, in order to avoid further processing of the results.

When considering the frame level ground truth, three different approaches can be adopted:
1. Only the ground truth corresponding to the central frame, or any given frame, of the analysis window is compared to the detector output for the given analysis window. This is the approach adopted in [2] and [3]. As discussed previously, this approach make obviously the problem very simple: there is little chance to fall on a silent frame. The ground truth is easily established and we have little chance to face the tricky situation where a speaker is starting or stopping to talk on that very frame.
2. The output of the detector for a given analysis window is compared to each of the t corresponding frames in the ground truth. By using this approach, the number of test points increases compared to the previous case. Thus the evaluation is more accurate. However, if the analysis windows are overlapping, some frames will be scored more than once in the final results. There might be contradictory labels at the output of the detector for a given frame, and additional processing steps are required.
3. A third, and more natural option in the case of overlapping analysis windows, might be to consider a trade-off between the two previous options, where a given fraction of the frame set belonging to the analysis window is compared to the output of the detector for the same set of frames. This last approach presents the advantage of being more accurate than the first one without the problem of multiple scoring frames, provided a judicious choice for the shifting window value and for the number of frames on which the algorithm performance is evaluated.

The advantage of using the frame level ground truth with approaches 2 or 3 rather than the window level ground truth is that it allows to cope with speaker turn points if these one fall in the current analysis window. Let us consider the case where the two persons are speaking about the same amount

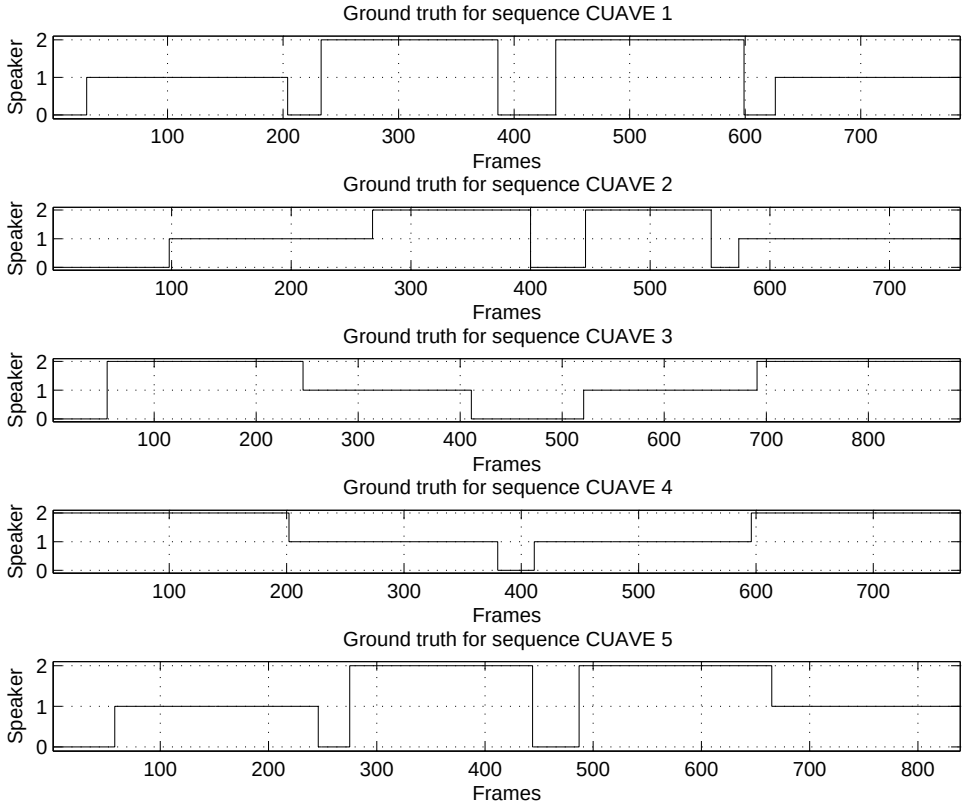

Fig. 2. Ground truth labels for the first five sequences of the group partition of the CUAVE database.

of time in a given analysis window where a turn point occurs. If the window level ground truth has been chosen, the output of the detector will be judged as either 100% right, or 100% false. Whereas, if the frame level ground truth has been chosen, the performance of the algorithm will anyway be proportionate to the situation.

In addition, the case of the silent frames must be carefully studied in both cases, and especially using the frame level ground truth. In this case, a "silent state" has to be considered when the number of consecutive silent frames is above a certain number L.

The last point to be discussed concerns the choice of the analysis window length and of the shift parameter value. The length of the analysis window has to be a trade-off between the algorithm requirements, the computation time, and the "inertia" of the detection (*i.e.* the delay between two detections). The value of the shift parameter determines the resolution of the detection algorithm. The best approach in that sense would be to shift the window by one frame all along the sequence. It is also the most expensive from a computation time point of view. The less time consuming approach would be, on the contrary, to make use of non-overlapping windows. But then the resolution is much lower and the results may drop significantly. In particular, this approach is not accurate enough to cope with the speaker turn points. Therefore a trade-off has to be found between accuracy and computation time.

V. Proposed experimental evaluation framework

Since our aim is that of evaluating speaker detection algorithms, in here we consider the multi-speaker partition of the CUAVE database. This section includes 22 sequences exhibiting two persons taking turn in reading series of digits. We have decided to build the ground truth using a frame level approach. Each frame has a label (0, 1 or 2), which indicates if no one (0), the left person (1) or the right person (2) is speaking. A group of frames is labeled as silent (0) when it is composed of at least

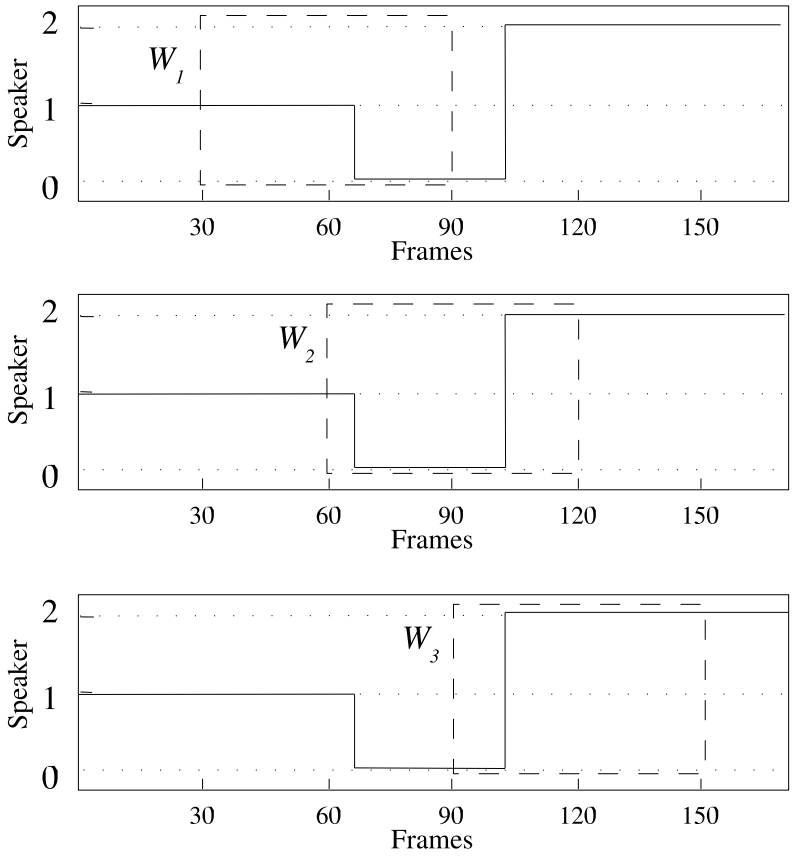

Fig. 3. Schematic representation of a sliding detection window applied on the ground truth. The detector output for each of the three window will be compared to the corresponding window ground truths: 1 for W_1 (top), 0 for W_2 (middle) and 2 for W_3 (bottom).

$L = 25$ frames. This value corresponds to the perceived limit between an interruption in the speech flow and a "true" silence. An example of the obtained labels for the first five sequences of the group partition of the CUAVE database is shown in Fig. 2. The complete set of labels for all the 22 sequences is available on the author's web page [4].

For what concerns the evaluation of the speaker detector's results, we propose to use a window-based evaluation method. Speaker detection algorithms typically output sets of frames denoted by a single speaker label. The detection is considered to be correct if the detector's output for a given window matches the most present label in the corresponding ground truth window (see Fig. 3).

VI. Conclusions

In this report, problems related to the evaluation of multimodal speaker detection methods have been discussed. An evaluation methodology has been proposed, in order to make possible the comparison between different algorithms, and the labeled ground truth for a multi-speaker audiovisual database, the CUAVE corpus, has been made available.

References

[1] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving-talker speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1189–1201, 2002.

[2] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proceedings of the International Conference on Image and Video Retrivial (CIVR)*, Urbana, IL, USA, July 2003, pp. 488–499.

[3] G. Monaci, O. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations [to appear]," *Signal Processing*, January 2006.

[4] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt, "Cuave database ground truth," January 2006. [Online]. Available: http://itswww.epfl .ch/∼besson/cuave_gt.html