

A Hybrid Approach to Online Speaker Diarization

Carlos Vaquero¹, Oriol Vinyals^{2,3}, Gerald Friedland³

¹University of Zaragoza, Zaragoza, Spain

²University of California, Berkeley, CA, USA

³International Computer Science Institute, Berkeley, CA, USA

cvaquero@unizar.es, {vinyals, fractor}@icsi.berkeley.edu

Abstract

This article presents a low-latency speaker diarization system (“who is speaking now?”) based on a hybrid approach that combines a traditional offline speaker diarization system (“who spoke when?”) with an online speaker identification system. The system fulfills all requirements of the diarization task, i.e. it does not need any a-priori information about the input, including no specific speaker models. After an initialization phase the approach allows a low-latency decision on the current speaker with an accuracy that is close to the underlying offline diarization system. The article describes the approach, evaluates the robustness of the system, and analyzes the latency/accuracy trade-off.

Index Terms: Speaker Diarization, online, incremental, hybrid

1. Introduction

Traditionally, the task of speaker diarization is to segment an audio signal into speaker-homogeneous regions addressing the question “who spoke when?” without any prior knowledge of the speakers, number of speakers, text, language, or amount of speech present in the recording [1]. As this definition implies, the task has mainly been addressed as an offline task. In other words, conventional systems make use of all available data in the recording before making a decision about how many speakers are present and when each of them is speaking. While offline processing offers the possibility to make use of long-term assumptions and optimize globally on the data, there are many applications, including dialog systems and videoconferencing, which require online processing or, informally, “who is speaking now?”. For example, a robot that interacts with several people might perform online diarization to turn its head to the actual speaker to make its response seem more natural. The major difficulty of online processing is that decisions are based on much less data. For example, at a given point in time, a speaker might enter the conversation who had not yet been registered by the system. A system that overcomes this problem using speaker identification with pre-trained speakers-specific models (as in [2]) would not be considered a speaker diarization system, as diarization requires no recording-specific a-priori training.

This article presents a novel hybrid online/offline speaker diarization system. The system consists of an online component that makes decisions with low-latency and an offline component that uses a traditional diarization approach running in the background to take advantage of all available audio information up to the current time. The models created by the offline system are then used to update the models of the online system. Intuitively, the farther one progresses into the meeting, the higher the accuracy of the system since the offline system can generate better models. On the other hand, more data for the offline system also

means higher latency and thus lower robustness against unseen data. The article describes the approach, evaluates the robustness of the system, and analyzes the latency/accuracy trade-off.

2. Related Work

Speaker diarization has become a mainstream research area in speech processing, and systems have improved and evolved dramatically in the last decade thanks in part to the NIST Rich Transcription evaluations. Even though an experimental low-latency task was introduced in the RT’09 evaluations, speaker diarization research so far has mostly focused on improving offline diarization performance.

In [3] a framework based on multimodal information and Dynamic Bayesian Networks was proposed with the goal of creating an online speaker diarization system. Initial experiments using the framework were encouraging, but the experimental setup was very controlled and consisted of a tiny dataset. Online speaker identification systems were also presented in [2] and in [4]. The approach presented in [5] fuses video information with audio captured using a microphone array to perform not only online speaker identification but also localization. However, all approaches require prior supervised training to obtain a model for each speaker and have no online adaptation (e.g. neither can detect new speakers).

An approach that is close to our proposed system is the one presented in [6], which was later refined in [7]. The approach in [7] is able to detect new speakers in a meeting without any prior knowledge of the speakers using audio from a single microphone. However, the system relies completely on accurate detection of new speakers and on speaker models that are trained according to online decisions. This strategy leads to error accumulation. Our system solves this problem through the use of hybrid online/offline processing, making use of all available information to train speaker models and not relying completely on online decisions, thus avoiding error propagation.

3. System Description

Figure 1 presents an overview of the proposed online diarization system. The system consists of two subsystems: An offline subsystem that generates speaker labels with certain latency for all available data so far, and an online subsystem that uses the labels generated by the offline subsystem to update speaker models and assign audio segments to speakers with low latency. The following subsections explain the approach in detail.

3.1. Feature extraction

From the audio stream, we extract 19 Mel-frequency cepstral coefficients (MFCCs) computed every 10 ms over a 30 ms win-

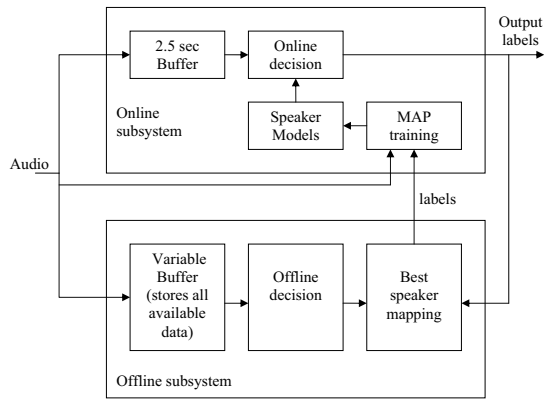


Figure 1: Block diagram of the hybrid online speaker diarization system as described in Section 3.

dow (with 20 ms overlap). This feature vector is the input for both offline and online subsystems. The offline subsystem models clusters with a Gaussian Mixture Model (GMM) of g Gaussians initially and increases as they are merged. The online subsystem makes use of a 128-Gaussian gender-independent Universal Background Model (UBM) trained on the AMI meeting corpus [11].

3.2. Offline subsystem

The offline subsystem uses all available data up to the current time to obtain speaker labels. Generally, the output of the offline subsystem is the more accurate in terms of Diarization Error Rate (DER), although the output is obtained with high latency. Therefore, the labels only serve to train speaker models. As shown in Figure 1, the offline subsystem first stores all available data in a buffer up to a time T_1 , and then the offline speaker diarization system obtains the labels for the segment $0 \dots T_1$. This output, consisting of a set of time intervals containing speech and its corresponding speaker clusters, is then compared with the output of the online subsystem up to time T_1 , in order to reassign speaker labels to obtain the best speaker mapping between the previous outputs of the online diarization system and the newly diarized sub-meeting. This step is necessary to keep consistency since speaker labels may change — diarization is an unsupervised task and the cluster numbering will, in general, differ between runs of the system. Upon completion of a run of the offline diarization system, a new instance is launched to perform diarization on the longer segment $0 \dots T_2$ ($T_2 > T_1$), and the whole process is repeated. The core of the offline subsystem is the ICSI speaker diarization system [8].

3.3. Online subsystem

The online subsystem produces the actual output of the online diarization system. This system uses all available data and labels up to time T_i (provided by the offline system as explained in the previous section), and outputs the most likely speaker given by majority voting on a 2.5 seconds, non-overlapping segment. Since lack of data (especially at the beginning of a meeting) will hurt system performance, speaker models are trained using Maximum a Posteriori (MAP) adaptation on the UBM. As shown in Figure 1, the online subsystem stores in a buffer the last available 2.5 s segment of the session, assigning it to the most likely speaker in our speaker model pool. The speaker adaptation process and the decision process are detailed below.

3.3.1. MAP adaptation

In order to train speaker models for the online subsystem, the output labels from the offline subsystem are used to adapt a UBM trained previously on the AMI meeting corpus [11]. This strategy is very similar to that used for speaker verification [9]. It was first applied for speaker diarization to improve the quality of the final iterations of a speaker clustering algorithm [10].

3.3.2. Majority voting

The decision for every 2.5 second segment is determined using majority vote [2]. For this task, every feature vector of the current segment is assigned to one and only one speaker of all known speakers according to a maximum likelihood criterion. Once all feature vectors have been assigned, the online subsystem selects the speaker who had the majority number of feature vectors of the current segment. The label associated to that speaker is taken from the output of the online subsystem, and thus the output of the online diarization system for the current segment. We have generally observed better performance using majority voting as opposed to e.g. maximum likelihood. This effect can be explained by the fact that some speaker models may lack training data or have very low speaker purity (as measured by the offline diarization system), so that using the actual likelihood values may add artificial artifacts to our window based decisions.

3.4. Operation

The offline subsystem is constantly running as a background process. Every time it finishes and outputs new speaker diarization labels, it restarts using all previous available data plus the data that arrived while it was running. We set one minute as the warm up time, so during the first minute the system will not output any speaker label. After the first minute the offline system begins running. There is therefore a delay while the offline subsystem performs diarization on the first minute of data and the online subsystem adapts the UBM to the speaker models. If the offline system works at $0.5 \times$ real time, the initial training labels will be available for the online subsystem to generate models 90 seconds after the session started, and at that point the offline subsystem will re-run on the 90 seconds available. The first output label will be obtained for the interval between the second 90.0 and the second 92.5 (the adaptation time is negligible compared to offline diarization time). This procedure is repeated until the whole meeting is analyzed. Note that as the available data increases, the latency to retrain the models is higher, but the labels used will be more accurate.

After each run of the offline subsystem, the given output is compared with the outputs that the diarization system gave for the same period of time. The speaker labels of the offline output are reassigned to obtain the best matching with the previous decision of the online diarization system, keeping the speaker model pool consistent.

The online subsystem, as explained above, makes use of those speaker models trained using the most recent output of the offline subsystem to decide in real time who is speaking. The latency of the system is about 2.6 seconds, which is the segment length plus negligible processing time. Once new labels are available from the offline subsystem, a new speaker model is trained for every speaker label, removing previous speaker models but keeping the numbering of the models consistent (otherwise the labels would change every time a new segmentation is obtained from the offline system). This procedure minimizes error propagation when a wrong decision is made.

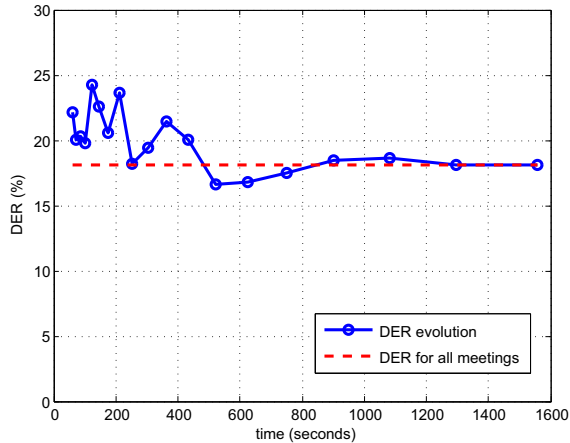


Figure 2: DER evolution of the offline subsystem against available data.

4. Evaluation

In this section we evaluate the performance of the online diarization system in terms of Diarization Error Rate (DER) and describe parameters effecting system performance.

4.1. Data and Configuration

For the following experiments, we took a subset of 26 meetings recorded in different environments with lengths between 5 and 20 minutes from the development data for the NIST Rich Transcription Evaluation 2009 (“Dev09”), removing all AMI meetings to ensure that no previous knowledge of the speakers was included (since the UBM was trained on AMI). Most speech activity detectors work incrementally and have accuracies in the high-ninety percents. Therefore, in order to isolate our results from hidden artifacts resulting from speech/non-speech error, we used ground truth speech/non-speech detection, keeping overlapped speech, so missed overlapping speakers count as missed speech when computing DER.

4.2. Offline diarization performance

The accuracy of the offline subsystem is critical for good performance of the overall system, since the speaker models that are used for the final online decision are trained on the output labels of the offline subsystem. In addition, speed of the offline subsystem is also important for two main reasons: First, we can expect our online system to work better as more data becomes available, since speaker models will be better estimated. For this purpose, the offline system should be fast, so as to make use of new data as soon as possible. Second, our proposed system has no separate method to detect new speakers. Once a new speaker appears in the session, the offline subsystem must collect data from the new speaker and perform a correct diarization on it before the online subsystem is able to detect the new speaker as a candidate. Therefore, if the offline system is slow, it may take a while to detect a new speaker — the faster the offline subsystem, the earlier the system detects a new speaker.

The measured offline subsystem performance improves as we increase the number of Gaussians. This improvement seems to stop at around 5 Gaussians (with a DER of 18.11%), and going up to 10 Gaussians gave little improvement. The computational cost increases as we increase the number of Gaussians, and it is thus desirable to minimize it, as it has a large impact

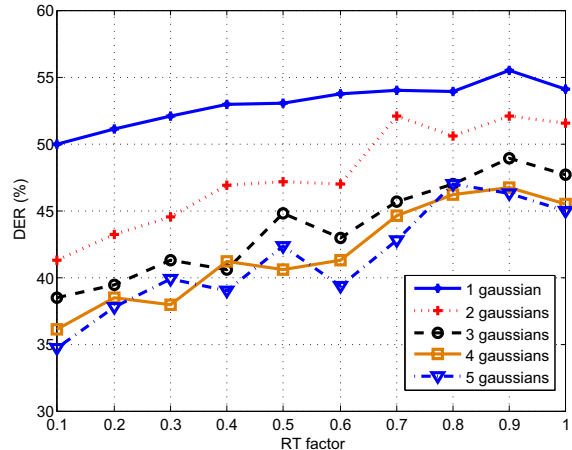


Figure 3: Overall DER as function of accuracy (number of Gaussians) and speed of the offline subsystem

on the online system as we describe later in this Section.

Next, we analyze how well the offline system does on shorter versions of the meeting (that is, considering only the first seconds in the meeting). The quality on those segments is crucial for achieving good DER. Setting the offline system to 5 Gaussians per initial GMM, which means using about $0.2 \times$ real-time to process the available data, and assuming that the warm up time is 60 seconds, the i th output labels of the offline subsystem will be obtained at the instant t :

$$t \approx 60(1 + RT\ factor)^i \quad (1)$$

In our case, $RT\ factor=0.2$, so the first output labels to train MAP speaker models obtained from the first 60 seconds of meeting will be available in $t=72$ seconds, the second output labels obtained from the first 72 seconds will be available in $t=86.4$ seconds, etc.

Figure 2 shows the evolution of the DER obtained for the 26 meetings by the offline subsystem as the available data increases (i.e. the DER by running and evaluating the system on the first seconds of the meetings). We observed that the DER decreases as more data is available converging to 18.11% when all the data is available. Even with only 60 seconds of input, the offline system performs well (22.35% DER); however, the models are undertrained, and, as we will see later, they are suboptimal if used on the whole meeting by the online subsystem (thus the need to keep updating models as the offline system produces new labels as data becomes available).

4.3. Online diarization performance

To study the online diarization performance, we analyze the DER obtained by the online subsystem (and thus by the online diarization system) depending on the performance and the speed of the offline subsystem. To obtain different performances with the online subsystem we vary the number of Gaussians in the initial cluster models from 1 to 5, and we simulate different system speeds.

As can be seen in Figure 3, the accuracy of the offline system is a key element in the performance of the online system. However, we can also observe the importance of speed on DER. Observe that online DER improves as the offline diarization system speed increases. This is due to two main factors: First, new speakers may appear during the session, and a faster system

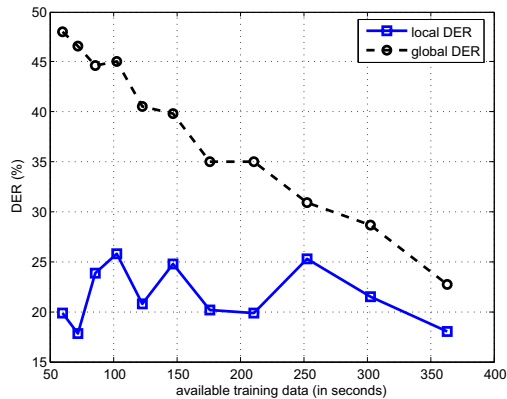


Figure 4: Evolution of the local and global DER (as defined in Section 4) as function of the amount of available training data.

will be able to obtain models faster, thus start outputting labels associated to new speakers earlier than a slower system. Second, known speaker models will be more accurate as there is more data available to train them. Therefore, a faster system will more frequently update the speaker models, obtaining better performance.

To show this effect, we can analyze the performance of the online subsystem depending on the data available for training. Assuming the same configuration again, we study the performance for different amounts of training data, analyzing local and global performance. Local performance is measured over the period of time the online system makes decisions with the given amount of available data (i.e. before the models are updated), while global performance is measured over the time remaining in the meeting, assuming we do not retrain any models. For example, for 60 seconds of available data, the local performance is measured analyzing the DER between 72 and 86.4 seconds, while the global performance will be measured from 72 seconds to the end of the meeting.

Figure 4 shows how the local and global DER evolves as the amount of available training data increases. The global DER behaves as we would expect: as the amount of available data increases, system performance improves. In contrast, performance for the local DER shows no clear pattern. This can be explained by the fact that, when there is little available data, the offline system re-processes the new data quickly, so the online system only makes use of the models obtained during a short period of time, which, in addition, is very close in time to the training data. For example, the first 60 seconds are used to train (poor) speaker models that are used from second 72 to second 86.4 (12 seconds after obtaining the data and for 14.4 seconds), but the first 250 seconds are used to train (better) speaker models that are used from second 300 to second 360 (50 seconds after obtaining the data and for 60 seconds). In the first case models are probably worse than in the second case, and the system has not seen all the speakers involved in the meeting, but the decisions that the system has to make with such models are easier than in the second case, where more speakers are known and better models have to decide among more speakers. Note that, despite the fact that the local DER fluctuates between 20% and 25%, the online DER is 37.75% as the local DER assumes optimal local mappings, while the online DER has to maintain a consistent mapping during every run of the offline subsystem, thus introducing more errors.

5. Conclusion and Future Work

In this paper we propose a hybrid low-latency speaker diarization system composed of offline diarization and online diarization subsystems. We demonstrate that it is important to obtain an accurate and fast offline diarization subsystem. Speed was shown to be almost as important as accuracy in obtaining good overall performance. We have analyzed the robustness of both the offline and online subsystems depending on the amount of available data. We showed that the offline subsystem performs very well even when not much data is available, which helps to obtain a good online diarization performance even at the start of a meeting, keeping the behavior of the overall system consistent. Our current methods allow building an offline diarization system that runs at $0.2\times$ realtime and obtains a 18.11% DER on a subset of Dev09 NIST Rich Transcription Evaluation set resulting in an overall online diarization performance of 37.75% DER. This result can be improved both increasing accuracy and speed of the offline diarization system, which we will primarily approach by parallelizing this part of the system. Fortunately, the proposed system is general as we do not rely on any particular offline diarization algorithm.

6. Acknowledgments

This research is supported by Microsoft (Award #024263) and Intel (Award #024894) funding and by matching funding by U.C. Discovery (Award #DIG07-10227).

7. References

- [1] Reynolds, D. A. and Torres-Carrasquillo, P., "Approaches and applications of audio diarization", In Proc. IEEE ICASSP, V:953–956, Philadelphia, PA, 2005.
- [2] Vinyals, O. and Friedland, G., "Towards semantic analysis of conversations: A system for the live identification of speakers in meetings". In Proc. IEEE ICSC'08, 426–431, Santa Clara, CA, 2008.
- [3] Noulas, A.K. and Krose, B.J.A., "On-line Multi-Modal speaker Diarization", International Conference on Multi-modal Interfaces, ICMI'07, 350–357, 2007.
- [4] Hung, H. and Friedland, G., "Towards Audio-Visual On-line Diarization of Participants In Group Meetings", Workshop on Multi-camera and multi-modal Sensor Fusion, M2SFA2, 2008.
- [5] Schmalenstroer, J. et al, "Fusing audio and Video Information for Online Speaker Diarization", in Proc. Interspeech'09, 1163–1166, Brighton, UK, 2009.
- [6] Markov, K. and Nakamura, S., "Never-Ending Learning System for On-line Speaker diarization", in Proc. IEEE ASRU'07, 699–704, Kyoto, Japan, 2007.
- [7] Markov, K. and Nakamura, S., "Improved Novelty detection for Online GMM based Speaker Diarization", in Proc. Interspeech'08, 363–366, Brisbane, Australia, 2008.
- [8] Wooters, C. and Huijbregts, M., "The ICSI RT07s speaker diarization system", in Proc. RT'07 Meeting Recognition Evaluation Workshop, 2007.
- [9] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Vol. 10(1-3): 19–41, 2000.
- [10] Zhu, X. et al, "Combining Speaker Identification and Bayesian Information Criterion for Speaker Diarization", in Proc. Interspeech'08, 2441–2444, Lisbon, Portugal, 2005.
- [11] AMI corpus, Online: <http://corpus.amiproject.org/>