Live speaker identification in meetings - "Who is speaking now?"

Oriol Vinyals and Gerald Friedland

TR-08-001

January, 2008

Abstract

The following paper presents an application that fuses the currently artificially separated tasks of speaker identification and speaker diarization. The presented method allows online identification of who is currently speaking using a single far-field microphone in a meeting scenario. It is able to recognize the current speaker after any two seconds of speech. An evaluation of the robustness of the algorithm using the AMI Meeting Corpus resulted in a Diarization Error Rate of 12.67 %.

LIVE SPEAKER IDENTIFICATION IN MEETINGS – "WHO IS SPEAKING NOW?"

Oriol Vinyals [1,2], *Gerald Friedland* [2]

[1] University of California, Berkeley, CA
[2] International Computer Science Institute, Berkeley, CA

ABSTRACT

The following paper presents an application that fuses the currently artificially separated tasks of speaker identification and speaker diarization. The presented method allows online identification of who is currently speaking using a single far-field microphone in a meeting scenario. It is able to recognize the current speaker after any two seconds of speech. An evaluation of the robustness of the algorithm using the AMI Meeting Corpus resulted in a Diarization Error Rate of 12.67 %.

Index Terms— application, speaker diarization, speaker identification, online

1. INTRODUCTION

Currently, speaker identification and speaker diarization are usually treated as two different tasks. The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question "who spoke when?" [1]. Many state-of-the-art systems use a combination of agglomerative clustering with Bayesian Information Criterion (BIC) [2] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [1]. These systems obtain satisfactory accuracy in terms of speaker diarization error. However, these approaches must process the entire file to perform the clustering. A system that is able to work incrementally on incoming data allows for a wide range of interactive applications. Past studies on the broadcast news domain proposed using a Universal Background Model (UBM) based system with real-time performance to detect speaker boundaries [3]. Another approach involved BIC and microphone array beamforming, which required detailed information on the location of the microphones [4].

The goal of speaker identification/verification is to detect a person's identity and possibly distinguish it from any possible impostors. In the classic speaker identification scenario, many utterances of speech from a large number of different people is given, and the task consists of mapping each utterance to each speaker. Common training sets consist of many hours of speech [5], and the test data must usually be several ten of seconds long; five seconds is considered a very short utterance.

In this paper, we present a novel application that falls in between the two conventionally separated communities. It performs online speaker identification, answering this question, "who is speaking now?", identifying the current speaker from a small set of speakers online, and in real-time. Using less than sixty seconds of training, the system is able to identify any of the known speakers after any two seconds of speech. The basic idea is to generate an environment model as well as speaker models and perform detection combining ideas from the classic GMM clustering and Speaker Identification. In addition, we developed a Graphical User Interface (GUI) to visualize and test our online system (Figure 1) in a real meeting scenario.

The rest of this article is organized as follows: Section 2 introduces the framework of our online diarization system; Section 3 shows the experiments and results; Section 4 finally summarizes this article and points out future work.

2. SYSTEM DESCRIPTION

For several years much effort has been made toward building robust speaker diarization systems. Though it is beyond the scope of this paper to review all the previous work done on the task, it is important to point out some important facts about many of the systems. In particular, we will be studying the ICSI diarization system [6] (referred hereafter as the "offline diarization system"), which uses a Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) based on 19-dimensional MFCC features to model speakers. The system starts with many clusters and, with a merging procedure based on Bayesian Information Criterion (BIC), guesses the total number of speakers and the temporal segmentation of a given meeting. This system performed very well on past NIST evaluations [6].

One of the limitations of the system described above is its computational cost. The running time for this algorithms is between three and five times the duration of the meeting. In a recent paper [7], a modification to the ICSI system that achieved

Fig. 1. Screenshot of the Java GUI that we developed to test the online diarization system.

below real-time performance without affecting accuracy was reported. Given that speaker diarization is much more efficient now, one can move forward to the next step: achieving online diarization.

We studied how some parts of the offline system could be simplified or replaced without a significant loss of performance. The use of the HMM could be replaced by a simpler procedure. Experiments on various meeting corpora suggested that one could use a temporal sliding window based on either maximum likelihood or majority voting to segment the audio data into chunks. For window sizes between one and two seconds the performance in terms of Diarization Error Rate (DER) was the same.

The new temporal window approach was compatible with an online system (HMMs, in contrast, require maximizing a global objective function). Assuming the speaker models are trained, the online system would have the same performance as the offline system. There are several options on how to build the models. If we want to keep the task of diarization unsupervised, we could run the offline diarization system on a meeting where everyone introduces himself, and use those models to segment future meetings. There is another approach that requires labeled data from the participants of a meeting to build a model for each participant. This is the solution that was adopted, and since it is a supervised approach, the system becomes a hybrid between speaker identification and diarization. Experiments and results of these two approaches are described further in Section 3.

The supervised system described above was built using a Java GUI interface (Figure 1) that was tested by non-expert users. The idea behind the interface is to have two different steps clearly identified: the enrollment, or training, step, where people have to sit in front of a computer and speak; and the segmentation step, where several people can speak while the system recognizes who is speaking on-the-fly.

2.1. Channel and unknown speaker issues

The online system presents two main challenging problems: channel robustness and dealing with unknown speakers. The first problem has been a focus of the speech community for a long time, and there is still a lot of effort to make features less channel dependent. Cepstral Mean Subtraction (CMS) was implemented to help deal with stationary channel effects. Although in subtracting the mean some speaker dependent information is lost [8], according to the experiments performed, the major part of the discriminant information remains in the temporal varying signal.

System	DEV07 DER	AMI DER
Baseline (ICSI system)	15.93 %	13.27 %
Baseline w/o HMM, 2.0s win	15.7 %	14.21 %
Baseline w/o HMM, 1.5s win	15.07 %	13.58 %
Baseline w/o HMM, 1.0s win	14.99 %	13.43 %
Baseline w/o HMM, 0.5s win	16.49 %	14.18 %

Table 1. Comparison between the baseline system and a modification using a window based segmentation with different windows lengths.

To address the second problem, we introduced several confidence measures to determine whether a speaker is in our speaker database or if it is a new speaker. The results and robustness of several measures are discussed in Section 3.

3. EXPERIMENTS AND RESULTS

The baseline system used in our experiments is the ICSI diarization system (offline system). Although it is not clear how to compare the two systems (the tasks are slightly different), we will use the DER scores of the offline system to depict what kind of values a state-of-the-art system can achieve. The two data sets of meetings used in this section are the development set for NIST RT07s Meeting Evaluation [6] which contains 21 meetings of past NIST evaluations (referred to as DEV07 in the rest of this paper), and a subset of 20 IS meetings of the AMI Corpus[1], recorded at IDIAP (referred to as AMI in the rest of this paper). The AMI meetings are a convenient choice since the 20 meetings are split into five different sessions, each one containing four meetings with the same four participants.

As explained in Section 2, the use of HMMs, which does not allow us to perform an online segmentation, is compared against a simpler windowing method. The DER on DEV07 and AMI data using a non-overlapping sliding window on frames of 10 ms is compared against the baseline system. For this experiment, a speaker in a given window was detected by majority voting on GMM likelihoods. This also achieves a faster detection since there is a potential saving in computing likelihoods (some likelihood computations can be skipped if someone already has 51 % of the votes). The maximum likelihood approach performed similarly.

The results of these experiments are shown in Table 1. The DER is very similar across all the systems. A significance test was run and no difference was observed between the systems with windows of 2, 1.5 or 1 seconds and the baseline. Note that choosing a bigger window implies more quantization error and choosing a smaller window implies more noise in the output. Given these considerations, a window between 1 and 2 seconds is a reasonable choice. Since using such a window does not affect the overall performance, the key question to answer is how to generate the speaker models with an approach that does not require all the data, as opposed to the agglomerative clustering that the baseline system performs.

3.1. Unsupervised online diarization approach

As mentioned in Section 2, there are two different approaches that can be used to build our models. The first approach is based on an inherent characteristic of the diarization task: being unsupervised. For this we require the data of a meeting where all the speakers participate. As a result, the offline system can be used to find the speaker models and use this models for other meetings where all the participants (or a subset of them) are present. For convenience, the AMI meetings were used to perform this test. One of the four meeting sessions is picked at random to run the offline system and the online system is used on the other three.

Table 2 shows the results using this approach. One important fact to note is that when the offline system fails (i.e., it does not find the right number of speakers, either because of deletion or insertion of speaker models), the performance of the online system that uses this models is significantly worse ("Worst trained DER" is 20.45 %). On the other hand, when the performance of the offline system is good enough, the performance is about the same as our baseline ("Best trained DER" is 14.19 %). This suggests that using pure data to build speaker models, as is proposed in the second approach, is feasible.

[1]The AMI Meeting Corpus, `http://corpus.amiproject.org/`

System	AMI DER
Baseline (ICSI system)	13.27%
Baseline w/o HMM, 2.0s win	14.21%
Best trained, 2.0s win	14.19%
Worst trained, 2.0s win	20.45%

Table 2. Comparison between the baseline system and an unsupervised approach that requires running the offline system on a meeting that contains the target speakers.

AMI DER	1 Gauss	5 Gauss	20 Gauss	100 Gauss
20 sec	25.38 %	16.56 %	15.80 %	15.93 %
50 sec	18.41 %	13.97 %	12.67 %	12.47 %
90 sec	17.60 %	13.11 %	12.50 %	11.99 %

Table 3. Results on how the training size and the number of gaussians to use to train the models affects the DER.

3.2. Supervised online diarization approach

In this section, we present experiments on how the models should be trained in order to achieve similar performance in terms of DER. The AMI data was chosen to run the experiments, as each of the five different scenarios contain approximately two hours of speech of four participants. One meeting is randomly chosen to train the four speaker models. The online procedure is used on the rest of the data to perform diarization. The amount of speech used (per speaker) as well as the number of gaussians used to train the models are shown in Table 3.

Surprisingly, with only 50 seconds of speech per speaker the system is able to perform the diarization task on all 20 meetings, a total length of more than 9.5 hours, with a DER better than the offline system. However, the fact that the DER is lower is to be expected: the online system has more information as it knows the number of speakers and has 50 seconds of speech for each of them. A final test was performed, consisting of building a bigger database of all the speakers present in the meetings (totaling 20 speakers) and then running the online system (20 gaussians and 50 seconds of training). This gave a slightly worse DER of 14.51 %, as each window could potentially be mapped to more speakers.

3.3. Unknown speaker detection

In order to determine if someone who is speaking is not in the database, the system needs information regarding the confidence level of a decision. Experiments were done using three different features:

- The ratio between the number of frames where the winning speaker had the highest likelihood and the total number of frames.
- The ratio between the likelihood of the winning speaker and the sum of the likelihoods of the rest of the speakers.
- The ratio between the likelihood of the winning speaker and the frame based sum of the rank-2 likelihoods.

All these features were effective in determining if a speaker is or is not in the database (lower values meaning less confidence), but the third one is less dependent on how many speakers are (making it less scenario dependent). Figure 2 shows the histogram of this feature.

3.4. The GUI based online system

A Java GUI application was built in order to further test the system in a real-life situation. Windows of two seconds were processed based on models trained with 50 seconds of speech and having 20 gaussians. Using a MacBook Pro and its built-in microphone, the process time of this two second window took about 300 ms, including the feature extraction step. Since performing Speech Activity Detection (SAD) using any existing system was an additional step that then became the bottleneck of the system, an additional model for non-speech was introduced, recording 60 seconds of several background sounds and silence.

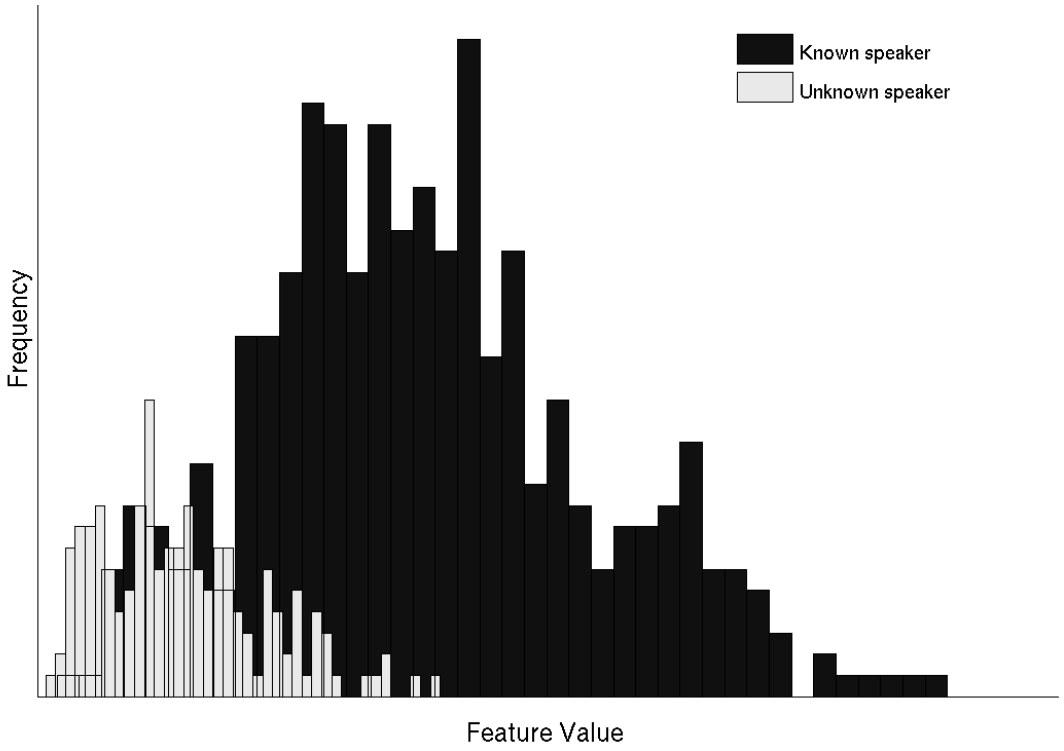

Fig. 2. Histogram of a feature to determine the confidence value of a decision. As can be seen, unknown speakers have lower confidence values.

This GUI was extensively tested by non-expert users, and it showed good performance, though we could not benchmark it due to the lack of ground truth annotation. We should note that in some cases the detection failed due to different channel conditions (especially when the models were not trained in the same room). Some unknown speaker tests performed satisfactorily, but a threshold has to be set depending on the application: there is a tradeoff between detecting people that are not in the database and missing people who are actually in the database.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to merge the classic task definitions of speaker identification and speaker diarization. The fused online speaker identification approach offers both a wide range of applications as well as a wide range of new challenges. With this new system it would be possible to analyze a meeting and adapt the environmental conditions of the room to who is speaking: changing the camera to point at who is speaking or changing the slide presentation on the whiteboard are some of the potential applications.

We performed a step by step simplification of the state-of-the-art ICSI diarization system to analyze how to build a robust online system. The experiments were tested using common meeting databases. A Java GUI was built for convenience to test how the online system worked in a real environment. All the results found in the databases were consistent with what was observed on the final system, achieving similar performance to the offline system.

However, further experiments on how to compensate for channel effects have yet to be explored in detail. Using some sort of Universal Background Model (UBM) could allow us to reduce the amount of time that a user has to speak to train the model, as well as introduce another way to deal with unknown speakers.

The final goal would be to achieve *real* online diarization (as diarization is inherently unsupervised), meaning that no previous data is needed to answer the question,“who is speaking now?”.

5. ACKNOWLEDGEMENTS

The authors would like to thank Adam Janin, Beatriz Trueba, Yan Huang, Kofi Boakye and Nelson Morgan for helpful comments on this paper and on the GUI application.

6. REFERENCES

[1] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE ICASSP*, 2005.

[2] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proceedings of DARPA speech recognition workshop*, 1998.

[3] TingYao Wu *et al.*, "UBM-based real-time speaker segmentation for broadcasting news," in *Proceedings of the IEEE ICASSP*, 2003.

[4] J. Schmalenstroeer and R. Haeb-Umbach, "Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming," in *Proceedings of Interspeech*, 2006.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE ICASSP*, 1992.

[6] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*.

[7] Y. Huang, O. Vinyals, G. Friedland, C. Mueller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *Proceedings of IEEE ASRU (to appear)*, 2007.

[8] Douglas A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.