Towards Semantic Analysis of Conversations: A System for the Live Identification of Speakers in Meetings

Oriol Vinyals and Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
{vinyals, fractor}@icsi.berkeley.edu

Abstract

In the following article we present an application that enables online identification of who is currently speaking using a single far-field microphone in a meeting scenario. By leveraging techniques from both the field of speaker identification and speaker diarization, the system 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 and the NIST Speaker Diarization Development set resulted in a Diarization Error Rate of 12.67 %.

Figure 1. The presented system at work: A laptop is not only recording the meeting but also identfying speakers as they talk. Only the laptop's internal microphone is used.

1 Introduction

In speech research, speaker identification and speaker diarization are currently 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?" [6]. The task is performed without prior training of specific speaker models. In fact, many systems work completely unsupervised, i.e. they do not require any a-priori knowledge. Current state-of-the-art systems, however, require the processing of an entire file and thus do not work online. Furthermore, since no prior knowledge is used, speaker diarization is not able to output real names. The goal of speaker identification/verification is to detect a person's identity and distinguish it from any possible impostors. In the classic speaker identification scenario, the test data usually needs to be about ten seconds long. Five seconds, an impossibly large latency for an online system, is considered a very short utterance.

In the following article, we present a system that merges techniques from the two different fields in order to create a novel online speaker identification system. It performs online speaker identification, answering this question, "who is speaking now?". It identifies the current speaker from a set of pre-trained speaker models 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 system has been implemented as a real application with a graphical user interface showing the name and the photo of the current speaker (see Figure 2). It is text and language independent. The performance of the system is tested on state-of-the-art benchmarks and compared against state-of-the-art speaker diarization approaches.

This paper is organized as follows: Section 2 starts by introducing the topic using previous work in the area before Section 3 introduces the system. Section 4 then presents some key experiments to verify the performance of the application on a larger scale. Section 5 dicusses the boundaries of the presented system before Section 6 finally concludes and points out future work.

978-0-7695-3279-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICSC.2008.58

Figure 2. Screenshot of the Java GUI that embeds the online speaker identification system. The system shows the face and the name of the current speaker, along with the time line (bottom) and the pool of trained speakers (right).

2 Related Work

Both robust speaker diarization and speaker identification are a active fields of research. It is beyond the scope of this paper to present a comprehensive review of the previous work. Please see [6] for a comprehensive introduction to the field.

Speaker diarization approaches can be organized into one-stage and two-stage algorithms, metric-based and probabilistic systems, and model-based and non-model-based systems.

Many state-of-the-art speaker diarization systems, including the ICSI Speaker Diarization engine (see Section 4), use a one-stage approach i.e., the combination of agglomerative clustering with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [2].

In two-stage speaker diarization approaches, the first step, speaker segmentation, aims at detecting speaker change points and is essentially a two-way classification/decision problem. In other words, at each point, a decision on whether this is a speaker change point or not needs to be made. After the speaker change detection, the speech segments, each of which contain only one speaker, are then clustered using either top-down or bottom-up clustering.

In model-based approaches, pre-trained speech and silence models are used for segmentation. The decision about speaker change is made based on frame assignment, i.e.

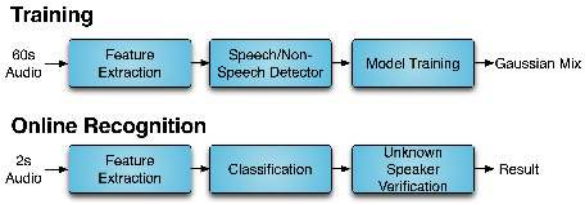

Figure 3. The main steps of the described system as outlined in Section 3.

the detected silence gaps are considered to be the speaker change points. Metric-based approaches are more often used for speaker segmentation. Usually, a metric between probabilistic models of two contiguous speech segments, such as Gaussian Mixture Models, is defined and the decision is made via a simple thresholding procedure. Over the years, research has concentrated on finding metrics for speaker change detection.

Other studies in 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 [7].

The task of speaker recognition is usually distinguished in two categories. In the first one, a speaker claims to be of a certain identity and his or her voice is used to verify this claim. This is usually called speaker verification. In speaker identification, the task is to determe an unknown speaker's identity. One application of speaker identification includes looking at a criminal's voice and cross-checking it against a database of criminal's voices looking for a match. Speaker recognition systems employ three styles of spoken input: text-dependent, text-prompted, and text-independent. Most text-independent speaker identification systems use a GMM/UBM [6] approach, along with a variety of channel normalization techniques [1] (e.g., feature-, model-, and score-level). Like speaker diarization, speaker identification is regularly evaluated by NIST. However, given their distinct field of application, speaker identification systems are tuned to require several seconds of input data. In the classic speaker identification scenario, many utterances of speech from a large number of different people are given, and the task consists of mapping each utterance to each speaker. Common training sets consist of many hours of speech [4], and the test data must usually be several tens of seconds long; five seconds is considered a very short utterance.

3 System Overview

The goal of the proposed system is to segment live-recorded audio into speaker-homogeneous regions with the goal of answering the question "who is speaking now?" .

In reality, however, the system has to actually answer the following questions:

- Is somebody currently talking?
- If yes: Is the speaker in the database?
- If yes: Who is it?

For the system to work online, the questions have to be answered on small chunks of the recorded audio data, and the decisions must not take longer than realtime.

Figure 3 shows a big-picture overview of the system. The steps are described as follows.

3.1 Training

In training mode, the user is asked to speak for one minute. The voice is recorded and converted to 19-dimensional MFCC features. A speech/non-speech detector is run. The speech segments are then used to train a Gaussian Mixture Model (GMM). The number of Guassians and iterations has been determined empirically, as described in Section 4.

In order to be able to cope with potentially difficult room conditions, e.g. air-conditioning noise, we also train an additional 60-second room-specific non-speech model.

3.2 Recognition

In the actual recognition mode, the system records and processes chunks of audio as follows. In a first step of feature extraction, the sampled audio data is converted into 19th-order MFCC features. Cepstral Mean Subtraction (CMS) is implemented to help deal with stationary channel effects. Although in subtracting the mean some speaker-dependent information is lost [5], according to the experiments performed, the major part of the discriminant information remains in the temporal varying signal.

In the classification step, the likelihood for each set of features is computed against each set of Gaussian Mixtures obtained in the training step. As determined by the experiments on larger meeting corpora (see Section 4), we use 2-second chunks of audio and a frame-length of 10 ms. This means, a total of 200 frames are examined to determine if an audio segment belongs to a certain speaker in the non-speech model. The decision is reached using majority vote on the likelihoods. If the audio segment is classified as speech, we compare the winning speaker model against the second best model by computing the likelihood ratio. This is a good indicator of the confidence level of the decision and enables the handling of unknown speakers. Figure 4 shows a typical histogram of this feature.

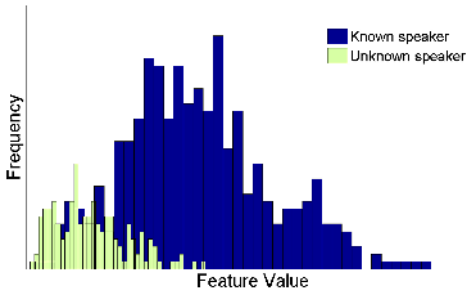

Figure 4. Histogram of the likelihood ratios (see Section 3) to determine the confidence value of a decision. In most cases, unknown speakers have lower confidence values as is shown in the figure.

All the above steps are computed on the fly, requiring less than 10 % real time using a Macbook Pro. The system has been embedded in a Java GUI. Figure 2 shows a screenshot. The Java application takes care of the recording of the meeting and stores the individual speaker segments as metadata. When a speaker is detected, the name and the face of the speaker are shown. An unknown speaker is visualized as a question mark, and non-speech is visualized as three dots.

4 Performance Evaluation

The following section presents a set of experiments to verify the performance and validity of the shown approach on a larger scale. Mainly, we justify that the transition we made from an offline system to an online system does not have an effect on accuracy. In fact, we show that the performance is comparable to state-of-the-art offline diarization approaches.

The baseline system used in our experiments is the ICSI diarization system (offline system). The ICSI Speaker Diarization System has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems[1]. The output of a speaker diarization system consists of metadata describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against

[1] Unfortunately, we are not allowed to present any ranking. Please refer to the NIST website for further information: http://www.nist.gov/speech/tests/rt/rt2007/

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.

manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate which is defined by NIST[2]. The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker-errors (mapped reference is not the same as hypothesized speaker). The current official score for the ICSI Diarization Engine is 21.74 % DER for the single-microphone case (RT07 evaluation set). This error can be decomposed in 6.8 % Speech/non Speech error and 14.9 % Speaker clustering error. The Speaker error includes all wrongly classified segments, including overlapped speech and very short segments.

The two data sets of meetings used in the experiments presented in this section are the development set for NIST RT07s Meeting Evaluation [8], 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[3], 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.

4.1 Optimal window length

State-of-the-art offline diarization systems rely on the use of HMMs in combination with the Viterbi algorithm. Such a global classification step is not possible in an online system as it would require the incorporation of the entire file. Therefore we have to limit ourselves to only a local classification. Since a frame-by-frame classification would be too noisy, we decided to use a sliding window of different sizes for a frame-by-frame majority vote (see Section 3). The following table shows how these results compare with the baseline system: the DER on DEV07 and AMI data using a non-overlapping sliding window on frames of 10 ms is compared against the baseline system. A speaker in a given window was detected by majority voting on the 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). As can be seen in Table 1, the maximum likelihood approach performed similarly as the DER is very similar across all the systems. Choosing a bigger window implies more quantization error and choosing a smaller window implies more noise in the output. Therefore a window size between 1 and 2 seconds is a reasonable choice.

[2]http://nist.gov/speech/tests/rt/rt2004/fall

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

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 2. Results on how the training size and the number of Gaussians to use to train the models affects the DER.

4.2 Supervised online diarization approach

In this section, we present the experiments that guided our choice on how the speaker models should be trained. The questions were: how many seconds of training data should be used to generate a speaker model and how many Gaussians should be trained? 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 meetings is randomly chosen to train the four speaker models. The online classification with a window length of 2 seconds 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 2. 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. Another test was performed that consisted 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.

4.3 Unsupervised online diarization approach

As a final experiment, we present an alternative model training approach. Instead of explicitly building speaker models by using 50 seconds of a certain speaker's speech, we used the output models of an offline diarization system as speaker models for the online diarization. This enables a silent-learning approach that does not require human attention. For example, the very first meeting could be diarized traditionally, and after this, all the meeting could reuse the speaker models for online speaker identification. Again, 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 3 shows the results using this method. Of course, 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 the 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 %).

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 3. Comparison between the baseline system and an unsupervised approach that trains the speaker models by running the offline system on a meeting that contains the target speakers.

5 Limits of the System

This application was tested by non-expert users, and seemed to work satisfactorily in most cases. Emotion such as laughter, coughs, or overlapped speech, are the cause for the majority of the recognition errors. Channel-dependency of models still poses a problem. Although both Cepstral Mean Subtraction and an adapted non-speech model help to improve the classification, in some cases speaker recognition failed because the models were trained in a different 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. This might limit the scalability of the system.

6 Conclusions and Future Work

In this paper, we presented a robust online speaker identification application for meetings. The system was derived from state-of-the-art research approaches and implemented as a usable GUI-based application. The performance was tested on two state-of-the-art meeting benchmarks and compares well with current research.

The semantic extracted by this novel meeting analysis approach allows for a wide range of applications, e.g. a system could automatically adapt environmental conditions of the room according to who is speaking, a camera view point could be automatically adjusted to point at who is speaking, or the right set of presentation slides could be loaded automatically when a new speaker starts talking at the podium.

The compensation of channel effects has yet to be explored more carefully. Using a Universal Background Model (UBM) might 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. A 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?".

7 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 application.

References

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrtaz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[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] T. W. *et al.* UBM-based real-time speaker segmentation for broadcasting news. In *Proceedings of the IEEE ICASSP*, 2003.

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

[5] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.

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

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

[8] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*, 2007.