Live Speaker Identification in Conversations

Gerald Friedland
International Computer Science Institute
Berkeley, CA
USA
fractor@icsi.berkeley.edu

Oriol Vinyals
University of California San Diego
San Diego, CA
USA
oriol18@gmail.com

ABSTRACT

The following article describes our technical demonstration of an online speaker identification system for conversations. A laptop with an internal microphone is centrally placed in the table of a meeting room. The system is able to identify the current speaker independent of spoken text or language with a latency of about 1.5 seconds and an accuracy of about 85 % (as evaluated against the NIST RT benchmark). A Java GUI shows the image of the current speaker along with a timeline containing past speakers. Speakers are added to the system's database using a one-minute training procedure.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing: indexing methods

General Terms

Algorithms, Applications

Keywords

speaker identification, diarization, conversations, online

1. INTRODUCTION

Currently, speaker identification and speaker diarization are treated as two different research fields. The goal of speaker diarization is to segment audio recorded using a single centralized microphone into speaker homogeneous regions. The goal is to answer the question "who spoke when?" [3]. 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. The output of such systems is therefore limited to labeling speaker regions with numbers or letters, but not with real names. Also, current state-of-the-art systems require the processing of entire files and thus do not work online. The goal of speaker identification is to detect a person's identity and reliably distinguish it from possible impostors. In the classic speaker identification scenario, the test data usually needs to be several ten seconds long. Five seconds, an impossibly large latency for an online system, is considered a very short utterance.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada
ACM 978-1-60558-303-7/08/10.

Figure 1: The demonstrated system at work: A laptop is recording the meeting and identifying speakers as they talk. Only the laptop's internal microphone is used and the system works text- and language independent.

Figure 1 shows the scenario setup of the presented system. People in a meeting are sitting around a lpatop and have a conversation. The laptop segments live-recorded audio into speaker-homogeneous regions with the goal of answering the question "who is speaking now?" and displays a picture associated with the currently active speaker. The following article shortly describes the technical background of the systems as well as its evaluation.

2. SYSTEM DESCRIPTION

In reality, answering the question "who is speaking now?", actually means answering "is somebody currently talking?", if yes, is the "speaker in the database?", and, if yes, "who is it?". For the system to perform live identification, the questions have to be answered on small chunks of the recorded audio data, and the decisions must not take longer than realtime. A big picture of the technical approach behind the system is sketched as follows.

In training mode, the user is asked to speak for 60 seconds. For the recognition to work properly, about 45 seconds of pure speech (i.e. no pauses) are needed. The voice is recorded and converted to 19-dimensional MFCC features, which have proved to be successful in speaker diarization [1]. We use a window size of 30 ms and a step size of 10 ms. Then, a speech/non-speech detector is run [5]. The speech segments are then concatenated and used to train a Gaussian Mixture Model (GMM). The number of Gaussians and iterations has been determined empirically, as described in [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 environment noise model.

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).

In recognition mode, the system constantly records audio and processes it as follows. Like in the training step, the sampled audio data (16 kHz, 16 bit, mono) is converted into MFCC features. Cepstral Mean Subtraction (CMS) is implemented to help deal with stationary channel effects [2]. This way, the system is less sensitive to varying channel conditions, such as reverberation changes due to changing numbers of people in the room. For every frame, the likelihood for each set of features is computed against each set of Gaussian Mixtures obtained in the training step, i.e. each speaker model and the non-speech model. A total of 150 frames is used for a majority vote on the likelihood values to determine the classification result. If the audio segment is classified as speech, we compare the winning speaker model against the second best model by computing the likelihood ratio. Since this is a good indicator of the confidence level of the decision, thresholding this value enables the detection of unknown speakers. On a Dual Core Mac Book Pro 2.0 GHz, a complete classification requires less than 10 % real time. Therefore the latency totals at about 1.5 seconds.

A Java GUI has been developed as a front-end to the system. It takes care of the recording in both training and recognition mode. When a speaker is detected, his or her name and associated photograph are shown. Figure 2 shows a screenshot.

2.1 Evaluation

In order to evaluate the robustness of the approach presented here, a series of experiments have been conducted that are described in detail in [4]. In order to compare the online approach against state-of-the-art (offline) speaker diarization, we compared the presented approach against the past development sets for the NIST RT meeting evaluation. We use a set of 21 meetings of all past NIST evaluations. The error is measured in terms of Diarization Error Rate (DER)

System	NIST RT DER
ICSI offline system	15.93 %
Presented online system	15.07 %

Table 1: Comparison between a state-of-the-art offline diarization system and the online system presented here.

as defined by NIST[1]. The Diarization Error Rate is composed of two components: speaker error (speech region, but wrong speaker selected) and speech/non-speech error (non-speech region classified as speaker or speech region classified as non-speech). Table 1 summarizes the results. Other experiments showed that the system is able to cope with a base of up to 20 speakers and still maintain it's accuracy.

3. CONCLUSION

We present a robust online speaker identification application for conversations, implemented with a demonstrable GUI interface. This application was tested by non-expert users and seemed to work satisfactorily in most cases. The underlying online identification system was tested using state-of-the-art meeting benchmarks and compares well with current research. Detecting the active speaker is very useful for a large range of applications, such as teleconferencing systems or as a preprocessing step to speaker-adaptive online speech recognition. Limits of the approach include sensitivity against incidental noise, channel variation, as well as laughter, coughs, or overlapped speech.

Acknowledgements

This research was partly funded by the German Academic Exchange Service (DAAD), IARPA VACE, and Fundaciòn Caja Madrid.

4. REFERENCES

[1] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Proceeding of the NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

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

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

[4] O. Vinyals and G. Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of IEEE International Conference on Semantic Computing (to appear)*, August 2008.

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

9[1]http://nist.gov/speech/tests/rt/rt2004/fall