Audiovisual fusion for speaker diarization

Noulas, A.

Publication date
2010

1

Introduction

Most people would not be impressed by a computer that synthesises speech or analyses short spoken segments. A scenario in which such skills are applied in a natural discussion, however, is still considered futuristic, because machines lack the ability to determine who spoke when, and who is addressed by the speaker. This problem, which requires the simultaneous analysis of audio and video information, is studied under the term speaker diarization and involves elements of Machine Learning, Signal Processing and Computer Vision.

Computer science research tackled the simplest version of this task under the term synchrony detection. In this case, the objective is to detect which region of the video modality appears most synchronised to the audio modality — this region does not necessarily contain a speaker, it could be the motion of a violin player and the corresponding melody. Taking this idea one step further, a short audiovisual segment can be examined to decide whether it contains a speaking person or not. In this case, the task is named speaker detection. Today, the widely accepted term speaker diarization corresponds to the task of segmenting a digital recording into speaker homogeneous parts and assigning each part to the corresponding speaker [107]. The segmentation part is often described as speaker change detection, while the assignment part is often treated as an independent task called speaker identification. In speaker diarization all the available information is used to determine the speaker at each part of the recording, e.g., voice or silence models, information from the rest of the recording, the location of the recording equipment, temporal information and so on.

Applications of speaker diarization

Humans perform speaker diarization subconsciously, and, consequently, its importance in a variety of tasks is often overlooked. Such tasks include Automatic Speech Recognition (ASR), automatic transcription, Human Computer Interaction (HCI) and human-to-human communication in video conferences.

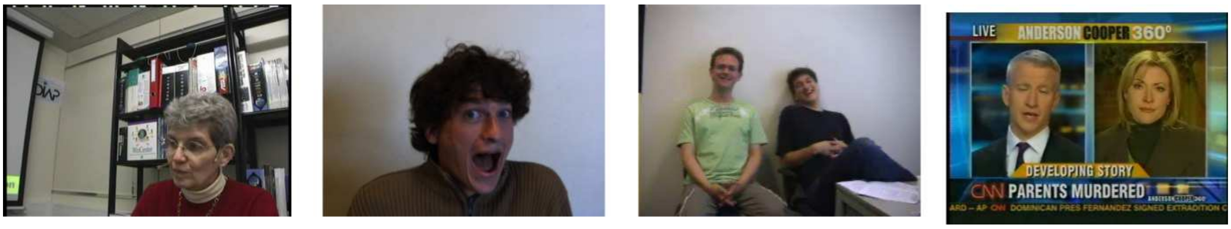

Figure 1.1: This thesis focuses on single audio, multiple camera recordings. The example frames shown here come from left to right from a smart meeting room, a video conference, an interview and a news broadcast recording.

ASR systems use speaker specific data to adapt a generic model before performing speech recognition. Speaker diarization can provide speaker-specific data in a novel recording, and, because of that, high accuracy speaker diarization is very beneficial for ASR [9]. Laskowski and Schultz further explore the effect of speaker diarization quality on ASR and Cuendet et al. investigate whether it is beneficial for ASR to use only speech parts with high confidence speaker diarization labels [37, 77].

Automatic transcription has been studied extensively in the National Institute of Standards and Technology (NIST) Rich Transcription (RT) evaluation benchmark. The goal of the evaluation is to produce transcriptions which are more readable by humans and more useful for machines [50]. These transcripts do not only contain the output of ASR, but also organise this output in terms of the speaker, and potentially augment it with meta-data describing the speaker's intention, focus or affects. In this evaluation, speaker diarization has been applied off-line to telephone speech [1], broadcast news [2, 75] and meeting videos [26].

HCI requires conversational agents that perform speaker diarization on-line and detect the speaker. Frameworks for on-line speaker diarization and the potential benefits of their use in HCI have been an active field of research for over ten years [38, 59]. One of the latest developments in the state-of-the-art robot ASIMO, involves performing basic speaker diarization [3] — identifying the source of sound, distinguishing between voice and other sounds and facing the people when being spoken to.

Human-to-human communication can be facilitated using speaker diarization in the context of video conferencing. Many people are visible and often the participants do not know each other in advance. A speaker diarization system can make it easier for the participants to follow the discussion, by providing additional visual feedback that indicates the speaker. Cutler et al. perform speaker diarization in a video conference and use their system to enhance human-to-human communication. They report qualitative results gathered by the participants' feedback, which demonstrate the added user-value of speaker diarization and indicate its potential applicability.

The objective of speaker diarization remains the same in all the tasks described above, but the quality and type of input modalities varies greatly. In genearal, a digital recording might contain modalities which facilitate speaker diarization in the form of meta-data or subtitles. A speaker diarization system, which will be applicable to most of the existing digital recordings, cannot assume the existence of such data and should limit the sources of information to the audio and video modalities. More specifically, this thesis focuses on audiovisual recordings containing a single audio track and one or more synchronised video streams. In the audio, this thesis makes no assumption about a fixed set of microphones or information about their position. In the video modality, one or more video inputs can be concatenated to a single video modality without any explicit knowledge of the camera locations, as long as the same person does not appear twice in any of the frames.

The choice of input modalities makes the methods proposed in this thesis applicable to most of the digital recordings existing today, ranging from web-camera videos to movies and smart meeting room sessions. Furthermore, most of the other recordings can be cast to such an input, by for instance merging the stereo audio input to a single channel. Such merging leads to information loss, but still allows for high-accuracy speaker diarization. Example frames of such recordings are shown in figure 1.1.

Issues in speaker diarization

Audiovisual data create a multimodal and high-dimensional input space, which makes speaker diarization challenging in many perspectives. A speaker diarization system has to deal with the following issues:

- The audio stream must be analysed, in order to detect the parts containing speech and model the voice of each person in a robust and reliable manner. Audio analysis has been a very active field of research in the signal processing and machine learning communities and is typically divided in two steps: extracting features from the raw signal input and using some statistical model to represent their distribution. Feature extraction is necessary because the original signal contains substantial information which proves useless and redundant in the context of speaker diarization. Moreover, the sampling rates for the audio modality are very high ranging from 16 kHz to 44kHz. A single sample contains no information, and therefore features are extracted from a larger part of the original signal in the form of a sliding window. The length of this window affects the results of speech analysis and should be selected carefully. The features extracted from the audio modality are noisy, and a statistical model is required to represent the information they provide. The voice of a person exhibits variation over time and the ideal model will be generic enough to capture this variation, and distinctive enough to distinguish each voice from the others. Note here, that similar processing is required for ASR or voice identification. The features used in speaker diarization tasks are closer to those of voice identification — the same phoneme coming from different speakers should be distinguishable.

- The video stream must be analysed, in order to detect the position of different persons and extract visual features which indicate speaking activity. The fact that humans spend 27% of their cognition abilities to perform vision, in contrast to just 7% for their hearing, demonstrates the complexity of vision in general [49]. Fortunately, the objectives of speaker diarization are much more constrained than those of general vision, and problem-specific techniques can be used. The first task is to perform face detection, in order to isolate the parts of the video stream that are potentially related to the speaker. Face detection is very robust since 2001 with the work of Viola and Jones [127]. These parts are still extremely high-dimensional and further processing is required to extract informative features for the task at hand, i.e., features that (1) are invariant to the variations in the appearance of a face because of illumination or facial differences in expression, and (2) distinguish one person from the other. Finally, a statistical model of the distribution of these features is necessary to deal with noise, occlusions and the temporal dynamics appearing in the face of a speaking person.

- Information coming from the audio and video modalities must be fused. On conceptual level, single modality analysis can provide information about the identity of the speaker, but this high-level information must be combined. Psychophysical research in humans shows that speaker diarization is influenced by both their hearing and sight. When humans localise sounds the phenomenon is known as the ventriloquism effect, and it is evident when a person is watching a movie and feels that the sound is coming from the lips of the actors rather than the speakers of the TV set [19]. When humans perform audio analysis the phenomenon is termed McGurk interference [83]. A common example is audiovisual recordings that are badly synchronised — the misalignment is easily detected but still hinders understanding. On signal level, there is useful information in the synchrony between the video and audio streams which are generated during speech. An audio and a video stream exhibit synchrony, when, at each point of time, the events existing in both streams have occurred simultaneously. Humans are very good in detecting such synchrony and use it to decide whether a person is speaking or not, and to associate the voice of a person to their appearance. In machine learning research, synchrony between the audio and video modality has been studied as a more general problem, including for instance the synchrony between the motion of a violin player and the corresponding violin melody. Speaker diarization requires to model the distribution of synchrony in the audiovisual feature space, and use it to perform synchrony and speaker detection.

- Speaker diarization is applied to audiovisual streams which are examples of sequential data. The complexity of sequential data analysis grows exponentially to the width and length of the sequence, which in this case corresponds to the number of appearing persons and the frames of the audiovisual recording respectively. Consider the example of a recording of three persons over 100 frames. If we represent with a binary variable the event of each person speaking on each frame, we will end up with a vector of length 300 which can take 2^{300} different values — only one of which corresponds to the correct speaker diarization output. In order to deal with this complexity, speaker diarization research makes assumptions regarding the width and temporal dimension of the recording. In the previous example, assuming that there is only one speaker at each frame, the width of the search space is decreased from $2^3 = 8$ to 3. In the temporal dimension, under a Markov assumption that the speaker at each frame depends only on the speaker at the previous frame, there exist dynamic programming techniques to estimate the optimal speaker diarization output [106].

These issues are far from solved, and current speaker diarization research focuses on remaining challenges in each one of these four categories. In the audio analysis one of the main challenges is not only to develop better models for the voice of different speakers, but to further combine them to predict the model created when two or more persons speak simultaneously [50]. Video analysis seeks algorithms that better track the visual features of a human face and visual features that can differentiate between spoken phonemes [135]. Modality fusion receives growing attention by researchers who try to deal with the different sampling rates of the audio and video streams, and their high dimensionality [17]. Finally, modelling sequential data is an open problem that has been studied extensively in the Machine Learning community [95].

The focus of this thesis is (1) the development of probabilistic methods for audiovisual data fusion and (2) the application of these methods on speaker diarization. The methods discussed in the following chapters make use of the state-of-the-art developments in single modality analysis, and focus on the challenges regarding modality fusion and sequential data modelling. The proposed methods remain, however, generic, in the sense that they can intuitively incorporate the output of future developments in audio and video analysis.

Thesis Position & Research Questions

Earlier computer vision, signal processing and machine learning research applied a variety of techniques on audiovisual data, ranging from neural networks [90] and rule-based systems [20] to fuzzy logic [25] and probabilistic models [132]. The position of this thesis is that the probabilistic modelling of audiovisual data is the first and foremost step towards robust speaker diarization and audiovisual data analysis in general. A probabilistic model of an audiovisual process has two main advantages. First, the complex parameters of audiovisual dynamics, such as a person's voice or appearance, can be learnt from unlabelled training data. Second, there exist well founded techniques to perform inference of the quantity in question from unseen examples, e.g., infer the identity of the speaker in a novel recording.

Specifically, this thesis addresses the following research questions:

1. What probabilistic framework can perform speaker diarization using information coming from the audio, video and audiovisual space?

2. How can we model synchrony between audio and video in speech and how can we use this to perform speaker diarization?

3. How can we model speaker detection using synchrony and how can we use the learnt models in speaker diarization?

The key component of these questions is the uncertainty associated with the quantities of interest: the identity of the speaker or the synchrony between the audio and video stream can not be measured directly. In contrast, these quantities should be inferred using the available audiovisual data. Relevant research, which is presented in detail in chapter 2, avoids directly modelling the complex multimodal dynamics of audiovisual data. For instance, speaker diarization is often performed using the audio modality alone [50, 132], while speaker and synchrony detection are often performed using simple heuristics [12, 59].

The choice of this thesis for speaker diarization is to incorporate the audiovisual data and the quantities in question under a probabilistic Bayesian framework, in the form of observed and latent random variables respectively. Bayesian approaches, presented in detail in chapter 3, have gained increasing popularity due to their capacity to deal with the uncertainty in the world perceived by a machine. In a nutshell, they provide an intuitive framework to incorporate prior knowledge about the domain (e.g., the number of persons in a recording), examine the observations at hand (audio or video features), and learn the model parameters from data. Moreover, provided the learning output and novel data, there exist inference algorithms to infer the state of the unknown variables, i.e., the identity of the speaker.

This thesis proposes Deep Belief Networks for synchrony detection between the audio and video in speech. Deep Belief Networks implement a product of simple distributions, which has the potential to model highly-varying high-dimensional distributions. This allows to (1) directly model the distribution of audiovisual features that reflect synchrony between audio and video and (2) intuitively apply this distribution for speaker detection.

Thesis Overview

Chapter 2 presents a review of the field of speaker diarization, which is divided in three parts: audio-based speaker diarization, synchrony-based speaker diarization and localisation-based speaker diarization. Initially, speaker diarization was tackled as an audio-based challenge and it was split in two subparts, speaker change detection and speaker identification. Chapter 2 organises the audio-based speaker diarization approaches in terms of their choices in these subparts. The NIST established a speaker diarization task in the RT evaluation benchmark in 2003, where the results of all approaches can be compared on publicly available data sets. The review describes in detail the method proposed by Wooters et al. [132], which is the winner of the 2007 evaluation [50] and at the time of this writing it is considered the state-of-the-art audio-based speaker diarization system. The review further covers a different line of research that performs speaker diarization indirectly through synchrony detection. In this line, researchers focused on detecting the location, over a sequence of frames, that appears most synchronised to the corresponding audio stream. In the case of speech, the detected location is the active speaker. The third part presents recent research on speaker diarization, which is based on audiovisual fusion. The review describes the frameworks that were proposed to combine information coming from the audio and video modality.

Chapter 3 focuses on the fusion of audio, visual and audiovisual information to perform speaker diarization in multispeaker recordings. The contribution of this chapter is a Dynamic Bayesian Network that performs speaker diarization fusing information coming from the audio, video and joint audiovisual space. The proposed Dynamic Bayesian Network implements a factorised transition model which resembles a factorial Hidden Markov Model (fHMM). It can be applied on a multispeaker recording and take into consideration the corresponding multispeaker relationships and the temporal dynamics of the process. The proposed model is tested in meeting and news videos, and improves the results in speaker diarization over the unimodal state-of-the-art framework.

Chapter 4 introduces an important contribution of this thesis, the application of Deep Belief Networks to capture audiovisual synchrony in speech. The framework of Deep Belief Networks is described in detail and the theoretical advantages of this approach are discussed. The proposed approach is tested on publicly available data sets and it is compared to synchrony detection methods proposed in relevant research.

Chapter 5 discusses how the detected synchrony can be used for speaker detection. The generative model represented by the Deep Belief Network is turned into a discriminative sigmoidal neural network, which is optimised to perform speaker detection. Chapter 5 compares the performance of the proposed model to that of models proposed in relevant research, and examines the suitability for the proposed neural network to the framework of Chapter 3.

Chapter 6 summarises the conclusions drawn from this thesis and discusses future directions that look most promising towards human-like automatic speaker diarization and high quality audiovisual information fusion.