

File ID 174200  
Filename 2: A review of speaker diarization

---

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation  
Title Audiovisual fusion for speaker diarization  
Author A. Noulas  
Faculty Faculty of Science  
Year 2010  
Pages vii, 167  
ISBN 90-75691-06-8

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/340830>

---

*Copyright*

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use.*

---

---

# A REVIEW OF SPEAKER DIARIZATION

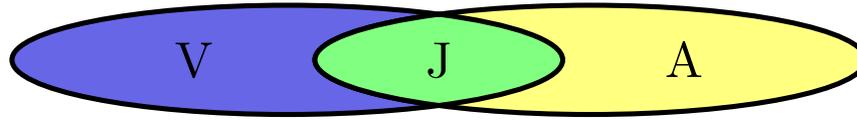
---

**Abstract:** This chapter reviews relevant research on speaker diarization. Speaker diarization incorporates advances in different scientific fields, namely signal processing, computer vision and machine learning. Consequently, some approaches to speaker diarization are barely comparable to each other. Section 2.1 gives an overview of the field, the different modalities used and the intuitive categorisation employed by this review. Section 2.2 contains the first category of speaker diarization approaches, which rely on information coming from the audio modality alone. Section 2.3 reviews synchrony-based speaker diarization approaches. These approaches measure synchrony between the audio and video stream, and assume that the person appearing most synchronised to the audio stream is the speaker. Section 2.4 presents localisation based approaches. Such approaches analyse the video modality to locate the position of the potential speakers, and the audio modality to locate the source of the audio stream. They assume that the person closer to the source of the audio stream is the speaker.

## 2.1 Overview

Speaker diarization, which is today a common term in the speech processing community, was introduced in the National Institute of Standards and Technology (NIST) Rich Transcription (RT) evaluation benchmark of 2003 as «Speaker Diarization - “Who spoke when”». It replaced the previous tasks of “speaker change detection” and “excerpt matching”. “Who spoke when” summarised the objective of speaker diarization, which is to cluster the audio stream in speaker homogeneous parts.

The creation of speaker diarization systems, like any automated procedure, is influenced by two factors. The way humans perform this task and the domain the system will be applicable to. Humans use their sight and hearing when following a multi-speaker interaction — they can recognise the face and voice of different people, and locate the source of their hearing input. Furthermore, they can detect synchrony between the audio and video modality to decide whether a person they see is speaking or not. Last but not least,



**Figure 2.1:** The different inputs used for speaker diarization. A stands for the audio modality, V for the video modality and J for the joint audiovisual space.

they use linguistic information which helps them anticipate when a person will continue speaking or stay silent. The domain of application for automatic speaker diarization is the existing digital recordings, e.g., smart meeting rooms sessions, movies, video conferences, phone conversations and news broadcasts — data available in large quantities, and for which accurate transcription of the spoken segments has many uses, e.g., retrieval, indexing or further processing.

This review is organised in terms of the system input which is illustrated in figure 2.1. The A space corresponds to the audio input, where, for example, the voice of the speaker can be identified. The V space corresponds to the video information, for example an estimate of the location of a potential speaker. The J space corresponds to information coming from the joint audiovisual observations. It could be information regarding whether the motion of a person’s lips is in agreement with the phonemes of the audio space.

The optimal system would use all the available information — even information that humans can not incorporate. That is, all the available video streams, all the available microphones, prior information about the location of these devices and their calibration, as well as the linguistic semantics of the spoken segments. Moreover, if one or more of these sources of information were missing, the optimal system would perform speaker diarization based on the remaining information. Clearly, such a system would be extremely complex and relevant research has made various compromises. For example, humans can perform speaker diarization, even when they do not understand the spoken language. Thus, the automatic systems could perform very well without any linguistic information.

More specifically, this review organises the speaker diarization methods to three categories: audio-based, synchrony-based and localisation-based speaker diarization.

Audio-based systems use the audio modality alone — the A space. This choice is motivated by the fact that (1) speaker diarization is not meaningful without the audio modality and (2) audio-based speaker diarization is applicable to a wider range of scenarios than an audiovisual system. For example, in data sets of phone conversations there is no video modality, and in many recordings (such as documentaries) the speaker is not visible.

Systems of synchrony-based speaker diarization are applicable to recordings where all the persons are visible. In such a recording, the system detects which participant appears most synchronised to the audio stream using observations in the J space. This choice is motivated by the fact that many recordings available nowadays contain an audio stream and the corresponding synchronised video stream, and the speaker remains visible throughout the

recording. Furthermore, such a system would be directly transferable to Human Computer Interaction (HCI) where the speaker is standing in front of the machine.

Localisation-based systems analyse the input of one or more cameras to acquire the location of the potential speakers, and the input of multiple microphones to locate the source of the audio stream. The person closest to the source of the audio stream is selected as the speaker. Localisation-based speaker diarization systems use the  $A$  and  $V$  space, but not the  $J$  space.

## 2.2 Audio-based Speaker Diarization

The structure of the section for the audio-based speaker diarization methods follows the structure proposed in the thesis of Anguera [89], which lead to the development of the current state-of-the-art audio-based speaker diarization system [132].

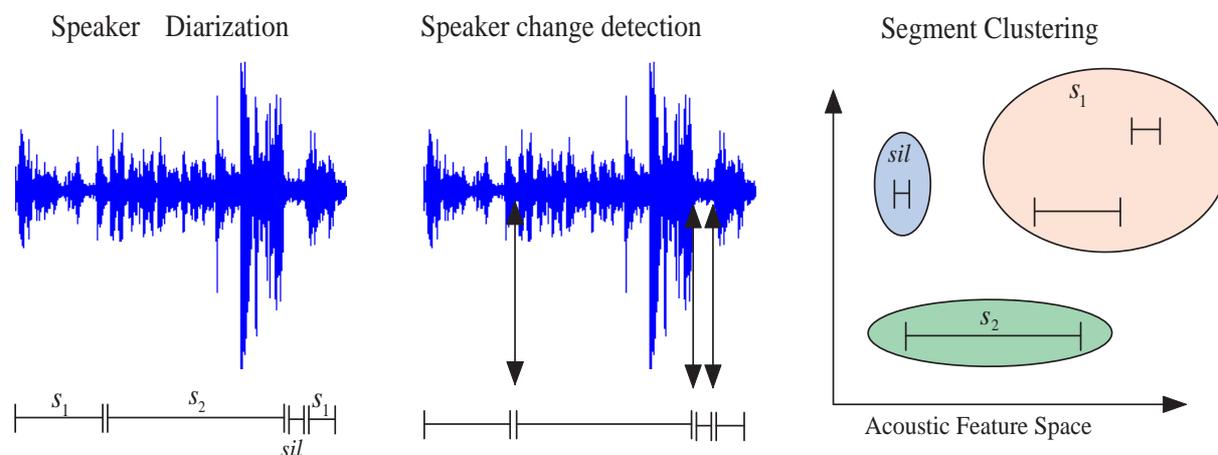
A visualisation of audio-based speaker diarization can be seen in figure 2.2. The audio input is noisy and it is sampled at a very high rate. Individual samples provide too little information to identify the correct speaker. The first step is to extract informative acoustic features from the raw input, a procedure described in detail in section 2.2.1. These features are extracted over a sliding window<sup>1</sup> that spans multiple audio samples. In principle, inference is made for each one of the extracted windows and their size is a very important choice.

On the one hand, long windows contain more information which makes it easier to select the correct label. However, the longer the window, the higher the chance that two persons speak in its duration, or a large part of the window corresponds to silence. In both of these cases, the window becomes less useful for Automatic Speech Recognition (ASR) and automatic transcription tasks. On the other hand, short windows contain very little information and are harder to classify. High resolution labelling, however, produces a much more beneficial output. This trade-off between precision and classification accuracy has given rise to different windows ranging from 0.25 to 4 seconds [4, 8, 11, 50].

The most typical approach is to break down speaker diarization in two parts, speaker change detection and speaker identification. Speaker change detection, which described in section 2.2.2, recovers the locations of transition between speakers. Speaker identification considers the long segments between the transitions as speaker homogeneous and labels them with the corresponding person's identity. Section 2.2.3 presents previous approaches to speaker identification. Section 2.2.4 presents speaker diarization as a NIST RT evaluation task, along with the details of the method of Wooters et al. [132], which is considered the state-of-the-art audio-based speaker diarization system.

---

<sup>1</sup>Often consecutive windows have small overlap



**Figure 2.2:** On the left, a graphical representation of audio-based speaker diarization, where the raw audio signal is shown on top and the desired output is shown at the bottom. The output contains two speakers ( $s_1$  and  $s_2$ ) and silence ( $sil$ ). The process is usually split in two parts, speaker change detection, which is shown in the middle, and clustering of the different segments in the acoustic feature space, which is shown on the right hand side.

### 2.2.1 Acoustic Features for Speaker Diarization

Speaker diarization is a speech-processing technique. In the framework of speech-processing, there are many well-known acoustic features including Mel Frequency Cepstral Coefficient (MFCC), Linear frequency cepstral coefficients, Perceptual Linear Predictive Coding, Linear Predictive Coding and others [5, 89]. A major issue regarding such features is that they were developed mainly for ASR — meaning that they were optimised to convey information about the phonemes rather than about the identity of the speaker. Nevertheless, MFCCs are the most common acoustic feature choice for speaker diarization [89]. In contrast with ASR approaches, speaker diarization methods use a larger range of frequencies, because higher frequency coefficients incorporate speaker identity information [89].

The MFCC descriptor, which was introduced in 1980 in the work of Davis and Mermelstein [44], is extracted through a four step operation. The first step takes the Fourier transform of the relevant window, while the second step maps the obtained powers of spectrum to the mel scale. The mel scale is a scale of pitch distances which resembles the way the human ear perceives sound, and the number of coefficients selected defines the final size of the MFCC descriptor. The third step computes the logarithm of the power at each frequency band, while the fourth step takes the discrete cosine transform of these logarithms, as if they were a signal. The size of the window used to estimate a descriptor varies, but the usual choice is between 10 and 25 milliseconds of data. At the time of this writing the

MFCCs are the most commonly used features for speaker diarization [5, 50, 54, 89, 132].

Alternative features focusing on speaker diarization have been proposed. For instance, Yamaguchi et al. [133] use energy of the acoustic signal, pitch frequency, peak-frequency centroid and peak frequency bandwidth and on top of them three novel features: temporal feature stability of the power spectra, spectral shape and white noise similarities. These features prove better than MFCCs for speaker diarization in two sets of experiments: news and meeting videos. This set of experiments was limited, and, moreover, the evaluation procedure involved training the model in one set and testing on the other. Thus, the proposed features are not applicable to general speaker diarization, in which no labelled training data is assumed.

Pelecanos and Sridharan in 2001 [102] proposed feature warping techniques which change the probability distribution function of the MFCC features to a Gaussian shape. In short, feature warping conditions and conforms the individual feature streams so as to follow a target distribution — which in this case is the Gaussian. Feature wrapping exhibited results robust to the influence of background noises and other non-speech events, which hinder automatic speaker diarization. This idea was originally proposed in the domain of speaker verification, but it was later applied with significant success to speaker diarization by Sinha et al. [116], Zhu et al. [136] and most recently Gupta et al. [58].

Up to the time of this writing, there has been no extended experimental evaluation that clearly favours one acoustic feature over the other. The focus of this thesis is on the sound probabilistic modelling of speaker diarization, and the widely adopted MFCCs features are used. However, the methods presented in the rest of this thesis are generic and can incorporate any choice of acoustic features.

### 2.2.2 Speaker Change Detection

Speaker change detection corresponds to locating the points of the audio stream where there is a change from one speaker to another or from speech to non-speech data. It is hard to precisely label these changes, even manually, at the acoustic descriptor level — which is a window of a few milliseconds. However, such accuracy is not needed and researchers have therefore used more coarse discretisations of time labelling windows of 0.3 to 2.4 seconds.

The length of the labelling window defines the shortest possible speaker-homogeneous segment and, therefore, affects non-speech detection. Even between phonemes within a word, there exist small pauses in the audio modality, corresponding to the pausal structure of speech [71]. Obviously, humans do not perceive them as silence and speaker diarization should not report them as silence either. Thus, the output will be silence only for pauses larger than the smallest possible predefined segment.

Errors in speaker change detection are of two kinds: missed changes, and false detection.

Missing a speaker change affects the speaker diarization results negatively. This is clear, since segments coming from two different speakers are assigned to only one. In contrast, a false speaker change detection might not affect the results at all, if both segments are assigned to the same speaker. Therefore, there is a tendency towards algorithms that generate false positives in speaker change detection rather than false negatives.

Speaker change detection has received increasing attention as a study field of its own (i.e., the thesis of Ajmera [5]). In this review, the relevant research is split in three categories, namely energy-based, model-based and metric-based speaker change detection. The same categorisation was employed in the review of Kemp et al. [72].

### **Energy-Based**

The simplest solution to the problem of speaker change detection is based on analysis of the acoustic energy of the audio stream and was traditionally applied in ASR frameworks. Energy-based approaches assume that all changes occur on silence segments. Thus, detecting silence segments in the stream corresponds to detecting the potential speaker change points.

Energy based systems use an acoustic feature that represents the energy over a sliding window and detect a change at the windows where this feature takes locally minimal values. The locations of minimum values are considered silence, and potential positions for speaker change. The sliding window returns numerous potential speaker change points and a threshold is used to decide which ones to keep [97, 128]. Alternatively, additional features over adjacent windows can be used. For instance Siu et al. measure the variability between these windows [117].

### **Model-based**

A straightforward machine learning approach to speaker change detection would involve modelling a closed set of the different audio classes, e.g., speaker A-speaker B-silence, speech-silence, speech-silence-music. Under the assumption that a recording contains only the classes of the closed set, the class-models can be applied to it in order to detect the speaker change points. The common choice for such approaches is to apply Gaussian Mixture Models (GMMs) to MFCCs. GMMs are chosen because they are universal density approximators, i.e., they can model an arbitrary probability distribution function over the data, while MFCCs are preferred because of their popularity in ASR. The audio stream is then classified to the Maximum Likelihood (ML) assignments [10, 72, 75, 80, 110].

Model-based methods perform badly in recordings which are very different from those of their training set and many approaches acquire these models directly from the novel data, either with a bottom-up or with a top-down approach. Bottom-up approaches set a large number of dense speaker change points and eliminate change points until some criterion is

met [4, 6]. Top-down approaches reverse this procedure: the original stream is considered one segment and additional segments are added until the desired criterion is met [8, 84, 85]. The stopping criterion corresponds to measuring statistics of the current clustering, and stopping the merging/splitting of clusters when a user-defined threshold is met, e.g., the total number of speakers or minimum segment length is reached.

### Measure-based

The most commonly used speaker change detection algorithms belong to the measure-based category<sup>2</sup>. Measure-based approaches measure the difference between two consecutive segments of the audio stream, which is usually referred to as distance between the two segments. If this distance is above a threshold, a speaker change is detected between the two segments.

The distance between two audio segments can be measured in two ways, using a closed form expression, or comparing a likelihood-based measure. The closed-form approaches extract the sufficient statistics of the distribution of the samples in each segment and combine these statistics in a closed form expression to acquire a distance measurement. This procedure is relatively fast and proves robust when the two segments contain enough samples to accurately estimate the sufficient statistics of their distribution. However, the requirement for a closed form distance function limits the choice of sufficient statistics to parameters of very simple distributions.

The likelihood-based measures compare the likelihood of the data under different model hypotheses. The different hypotheses are (1) to use a different model for each one of the two segments or (2) to use a single model for both segments. The ratio between the likelihood of the two hypotheses corresponds to the distance of the two segments. The computation of likelihood-based measures is much slower, since it involves training of models and evaluation on the segments. It can potentially provide much better results, however, since there is no restriction to the parameterisation of the models which are evaluated: models used in a closed form solution can be used under a likelihood-ratio comparison, while the converse is not always possible.

The most commonly used closed-form measure in the speaker diarization literature is the computation of the Kullback Leibler (KL) Divergence. The most commonly used likelihood-based measures are the Bayesian Information Criterion (BIC) and its extension in the form of the cross Bayesian Information Criterion (XBIC), and the Generalised Likelihood Ratio (GLR):

- KL divergence is a statistical measure of the difference between two distributions. Let  $Q_1$  denote the distribution over data  $\mathbf{x}$  under parameters  $\theta_1$ , i.e.,  $Q_1 = p(\mathbf{x}; \theta_1)$ .

---

<sup>2</sup>These methods are often called *metric-based* in relevant speaker diarization research. However, most of them do not produce a metric-space in the mathematical, and, therefore, will be described as measures in this work.

Given any two distributions  $Q_1$  and  $Q_2$  defined over the same random variable  $\mathbf{x}$  the KL divergence is given by:

$$\text{KL}(Q_1||Q_2) = \sum_{\mathbf{x}} Q_1(\mathbf{x}) \log \frac{Q_1(\mathbf{x})}{Q_2(\mathbf{x})} \quad (2.1)$$

where the sum is replaced by an integral in case  $\mathbf{x}$  is continuous. Intuitively, KL divergence will merge the most similar segments, regardless of how well they are modelled jointly or independently.

The advantage of KL divergence is the fact that for specific parametric forms of the distributions  $Q_1$  and  $Q_2$ , e.g., Gaussian, its value can be evaluated in closed form, while for others, e.g., GMM, there exist approximation techniques which are both fast and accurate. A disadvantage is the fact that KL divergence is not symmetric, that is,  $\text{KL}(Q_1||Q_2) \neq \text{KL}(Q_2||Q_1)$ , and therefore it is not a proper metric. A symmetric version of the KL divergence was proposed by Kullback and Leibler themselves, in the form of  $\text{KL}(Q_1||Q_2) + \text{KL}(Q_2||Q_1)$  [76].

The KL divergence was first used for speaker change detection in the work of Sigeler et al. [115]. In the work of Delacourt and Wellekens [45] the symmetric KL divergence was used as a part of a two step speaker change detection process, and their speaker change detection system based on the symmetric KL divergence was further improved in the work of Zochova and Vlasta [104]. Finally, a comparison of KL divergence versus the Mahalanobis and Bhattacharyya distance measures in terms of accuracy in speaker change detection was made in the work of Huang et al. [64].

- The Bayesian Information criterion was first introduced by Schwarz as a statistical measure to facilitate model choices [111]. The original formulation of BIC is:

$$\text{BIC}(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) - \frac{1}{2}|\boldsymbol{\theta}| \log(N) \quad (2.2)$$

where  $|\boldsymbol{\theta}|$  is the number of free parameters in model  $\boldsymbol{\theta}$  and  $N$  the number of independent data points in data  $\mathbf{x}$ .

In the speaker diarization literature this has been slightly modified. The objective is to decide whether two segments were produced from the same speaker or not. The modified BIC value for a segment  $\mathbf{x}$  modelled by parameters  $\boldsymbol{\theta}$  is:

$$\text{BIC}(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) - \frac{1}{2}\lambda|\boldsymbol{\theta}| \log(N) \quad (2.3)$$

where  $\lambda$  is a user-defined parameter which is used to change the importance of the value of the second term.

In speaker change detection between segments  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , the first hypothesis is that the segments should be modelled independently under distributions described by parameters  $\boldsymbol{\theta}_i$  and  $\boldsymbol{\theta}_j$  respectively. The second hypothesis is that the two segments

should be modelled jointly as one segment  $\mathbf{x}$ , under a distribution described by parameters  $\boldsymbol{\theta}$ . In that case, the distance measure becomes:

$$\Delta\text{BIC}(\mathbf{x}_i; \mathbf{x}_j) = \text{BIC}(\mathbf{x}, \boldsymbol{\theta}) - (\text{BIC}(\mathbf{x}_i, \boldsymbol{\theta}_i) + \text{BIC}(\mathbf{x}_j, \boldsymbol{\theta}_j)) \quad (2.4)$$

and the  $\lambda$  parameter is usually set such that equation 2.4 becomes positive when the two segments should be merged and negative when they should be modelled independently<sup>3</sup> — which corresponds to a speaker change detection.

The BIC criterion was first used for speaker change detection in the work of Chen and Gopalakrishnan [33, 34]. The first issue considering the application of the BIC criterion is to set the parameter  $\lambda$  automatically to a good value for a novel data set. Multiple approaches that used the BIC criterion proposed different ways to do that, and reported robustness to different audio contexts [45, 94, 124, 129].

The main issue with the BIC criterion, like all likelihood-based methods, is the computational cost. Obviously, it is computationally infeasible to compare all the possible segmentations to each other. Therefore Chen and Gopalakrishnan [33, 34], Tritschler and Gopinath [124], Cheng and Wang [114] and Cettolo and Vescovi [28, 126] propose a single or double pass. These passes are performed with growing windows that iteratively locate the speaker change points.

Anguera proposed the XBIC distance measure by cross likelihood evaluation [7]. The XBIC distance measure is inspired from the BIC criterion and the work of Juang and Rabiner in distance estimation between HMM [69]. It is computed as:

$$\text{XBIC}(\mathbf{x}_i; \mathbf{x}_j) = \log p(\mathbf{x}_i; \boldsymbol{\theta}_i) + \log p(\mathbf{x}_j; \boldsymbol{\theta}_j) \quad (2.5)$$

for each potential speaker change point, and the final speaker change points are selected by thresholding the computed XBIC value. XBIC distance is considered the state-of-the-art measure in speaker change detection [50].

- A second widely adopted distance measure is the Generalised Likelihood Ratio (GLR). GLR is a likelihood based measure that proposed to compare the two different model hypothesis through their ratio. This corresponds to:

$$\text{GLR}(\mathbf{x}_i; \mathbf{x}_j) = \frac{\log p(\mathbf{x}; \boldsymbol{\theta})}{\log p(\mathbf{x}_i; \boldsymbol{\theta}_i) + \log p(\mathbf{x}_j; \boldsymbol{\theta}_j)} \quad (2.6)$$

where the sum in the denominator among the log-probabilities corresponds to the multiplication of the actual probabilities.

Typically, the GLR is computed over two consecutive segments of fixed length and the candidate point is kept as a speaker change point if the GLR is above a threshold.

---

<sup>3</sup>Note here, that a threshold can be set directly on the value of  $\Delta\text{BIC}$  and the parameter  $\lambda$  can be avoided, but this is how the BIC is usually presented in speaker diarization literature.

GLR generally detects multiple speaker change candidate positions, and therefore its output is often processed further to eliminate erroneous detections. For example, in the work of Bonastre et al. [23] the threshold is set to a value that minimises the missed points at the cost of a higher rate of false detections. In the work of Gangadharaiah et al. [53] the GLR is used to segment the data and Viterbi decoding is then used to find the final speaker change points.

The best known speaker segmentation framework that relies on GLR is DISTBIC, introduced in the works of Delacourt et al. [45, 129]. DISTBIC performs a GLR pass that provides candidate change points and a second pass using the BIC criterion to select the final change points. The BIC criterion considers segments of non-fixed length and, consequently, it is slower but produces more robust results [129].

Multiple other measures have been suggested in the literature to perform speaker change detection, such as for example the Gish distance, introduced by Gish et al. [57]. All of them are based on the same principles of BIC, GLR or KL divergence, and no other measure has managed to get the broad acceptance of these three. A detailed review of all the previous approaches to speaker change detection is out of the scope of this thesis, and the interested reader is pointed to the work of Ajmera [5] and Anguera [7] that focus on speaker change detection and audio-based speaker diarization respectively.

## Conclusions

Speaker change detection is an important preprocessing step for audio-based speaker diarization. Accurate detection of all the speaker change points, with as few false positives as possible, is necessary to achieve high-accuracy speaker diarization results. Previous research is based on any of three key assumptions about the parameters of speaker change points: (1) that change points always coincide with silence, (2) that the features required to detect a speaker change point can be learnt off-line from training data, (3) that the distribution of observed features before and after the change point differ significantly.

Chapter 3 proposes a solution, which corresponds to a mixture of the second and third strategy. In a preprocessing step, training data is used to define a probability distribution over the number of speakers for each labelling window — in a recording containing three persons, this corresponds to the probability that no person (silence), one person, two persons or three persons are simultaneously speaking. The model then decodes the sequence using this distribution and parameters acquired from the novel recording, i.e., it performs speaker change detection and speaker identification in parallel.

### 2.2.3 Speaker Clustering

Speaker clustering is the task of assigning the audio segments created by speaker change detection to speaker homogeneous clusters. General clustering is a very well studied problem

in Machine Learning and the most common approaches to speaker clustering are applications of machine learning algorithms. This review divides the speaker clustering approaches proposed in the relevant research in two categories: hierarchical clustering and model-based clustering.

### Hierarchical Clustering

The hierarchical clustering is a simple and intuitive method to cluster the audio segments. Initially, the distance between each pair of speech segments is computed, using a user-defined distance measure which assigns smaller distances to acoustically similar segments. The clustering algorithm uses these distances to cluster the closest segments (in the bottom-up approach), or split the furthest away clusters (in the top-down approach). When the user-defined stopping criterion is met, the iterative procedure of clustering or splitting stops and the current clusters are the final detected speakers.

Clustering is usually performed in a bottom-up fashion. Chen et al., in [33, 34], were the first to propose bottom-up clustering using as distance measure the BIC criterion — see equation 2.3. In the work of Chen et al., the clusters are merged until there is no merging with a positive  $\Delta\text{BIC}$  — equation 2.4. Tritchler and Gopinath [124], Chen et al. in their later work of [30], Barras et al [11] and Cettolo and Viscovi [28] use a slightly modified version of the BIC criterion, by, e.g., adding a different penalty term in favour of specific clustering structures, evaluating windows of varying length, or using fast approximations to the computation of the BIC criterion. Gauvain et al. used bottom-up clustering and the GLR of two segments as distance measure [80].

Instead of using a distance measure, bottom-up clustering can also be performed under the assumption that the samples of the audio segments in a speaker-homogeneous cluster follow a specific parameasure distribution. Typically, this distribution is parameterised as a GMM in the acoustic feature space, which naturally leads to KL divergence as a distance measure between clusters. Unfortunately, there is no closed form for the KL divergence of two GMM and therefore various approximations are employed. For instance Beigi et al. [14] compares pairs of the closest components. The KL divergence between pairs of components can be easily computed — there exists a closed form expression for the KL divergence of two Gaussian distributions.

Ben et al. [16] and Moraru et al. [92] assume a GMM distribution with fixed variance and mixture proportions for each final speaker cluster. They compute the KL divergence for these restricted mixtures as

$$D(Q_1, Q_2) = \sqrt{\sum_i \sum_d \pi_i \frac{(\mu_1(i, d) - \mu_2(i, d))^2}{\sigma_i^2(i, d)}} \quad (2.7)$$

where  $\mathbf{i}$  represents a mixture component and  $\mathbf{d}$  a feature of the acoustic space, and use it as a distance measure between the distributions of two audio segments. In equation 2.7,

for each audio segment the variables  $\boldsymbol{\mu}$ ,  $\boldsymbol{\sigma}^2$  and  $\boldsymbol{\pi}$  contain the means, variances and prior probabilities of the components. The index  $(i, d)$  retrieves the  $d^{\text{th}}$  dimensional element of the mean  $\boldsymbol{\mu}$  or variance  $\boldsymbol{\sigma}^2$  of the  $i^{\text{th}}$  component and  $\pi_i$  retrieves its prior probability.

In both the works of Ben et al. and Moraru et al., the clusters with the smallest distance are merged iteratively, but the stopping criterion differs. Ben et al. use a threshold value to decide when to stop the iterative merging of clusters, while Moraru et al. use the BIC of the final model.

Top-down clustering is computationally much more expensive since all the possible splitting points should be compared. Therefore, it is the choice of far fewer researchers and it is applied with greedy approximations. For example, Johnson et al. [68] and Johnson [67] use top down clustering. At every iteration they first segment the current cluster in four subclusters, and then merge the most similar clusters again. Meignier et al. [84, 85] and Anguera and Hernando [8] follow a different line of hierarchical top-down clustering. In these works, the initial audio stream is used to adapt a cluster model and the clusters are split in order to optimise a likelihood measure of the stream; Meignier et al. split the clusters until the BIC of the final clustering reaches a threshold while Anguera and Hernando split the clusters until convergence of the final model likelihood.

## Model-based Clustering

The model-based clustering is typically performed in parallel to the segmentation. Multiple passes over the audio stream create the optimal segmentation, cluster the data, adjust the cluster model parameters and resegment the audio track. This procedure is iterated until convergence or until a stopping criterion is met. Examples of such speaker segmentation and clustering systems are the works of Ajmera et al. [6], Ajmera and Wooters [4] and Wooters et al. [131].

Ajmera et al. use a Hidden Markov Model (HMM)<sup>4</sup> with minimum duration constraints, i.e., the system must remain in a specific state for a minimum duration before transition to a different state. Clustering is performed iteratively, using the EM algorithm and Viterbi Decoding [6]. Ajmera and Wooters use a similar method (GMM and EM), but initialise the model with the k-means algorithm and stop the iterative procedure using the BIC criterion [4]. Wooters et al. detect and discard the non-speech segments and run Viterbi decoding over the remaining stream [132]. At the end of each iteration, they merge the clusters with the lowest  $\Delta\text{BIC}$  score, and rerun the Viterbi decoding. The procedure is continued until no cluster merging produces a positive  $\Delta\text{BIC}$  score.

---

<sup>4</sup>The HMM, EM algorithm and Viterbi decoding are described in detail in chapter 3

## Conclusions

The clustering methods used in audio-based speaker diarization come straight from the related machine learning research. The simplest hierarchical and model based clustering produces satisfactory results, and there has been no attempt to apply more complex clustering algorithms. This is because the errors coming from speaker clustering are due to the high variation in a person's voice rather than a direct outcome of the clustering procedure.

### 2.2.4 The NIST RT Evaluation

The variety of methods for speaker diarization and the importance of its output for high-quality ASR, led the leading ASR evaluation benchmark, the NIST RT, to include speaker diarization as an independent task, in the Speaker Diarization - "Who spoke when" evaluation benchmark of 2003. The objective of speaker diarization was changed slightly from one evaluation to the other, ranging from clustering of presegmented speech parts to a preprocessing step of each ASR system. Since 2005 though, it corresponds to segmenting a stream in speaker homogeneous parts, meaning automatically detecting the non-speech segments of the stream and assigning the speech parts to the corresponding speaker.

In the NIST RT-07 evaluation [50], seven different research groups participated in the speaker diarization task, all using the audio modality alone. The data was coming from smart meeting rooms, lecture recordings and coffee breaks, and the methods were evaluated on ten to twelve minute excerpts. During the evaluation of different methods, a 0.25 seconds "collar" is used for the human annotation, i.e., there is no evaluation of the labelling of points 0.25 seconds around a human-annotated speaker change point. This lowers reported speaker Diarization Error (DER), since high temporal precision in the annotation results is not necessary and short utterances, which are hard to detect and classify, are not evaluated. The best performance was exhibited by the framework submitted by Wooters and Huijbregts [132], which had a total DER of less than 15%.

### Practical implementation of the Wooters' speaker diarization system

The framework of Wooters and Huijbregts performs speaker diarization in two stages. First, speech segments are separated from non-speech segments. Then, hierarchical bottom-up clustering is performed in the speech segments to acquire the final speaker diarization results.

The speech/non-speech detector is a three step algorithm. In the first step, the audio stream is divided into speech and non-speech data, using an HMM which was trained on news broadcast data. The non-speech consists of silence and non-speech sounds, so that this region is further divided into two parts, one containing the low energy and one containing the high energy segments. The low energy non-speech segments are expected to

correspond to silence, and the high energy non-speech segments to non-speech sounds. The data that was originally classified as speech is used to train a GMM of 24 components. The low energy non-speech data, which are labelled as silence, are used to train a 7 component GMM. The high energy non-speech data, which are the non-speech sounds, are used for an 18 component GMM. The number of components for each class is set empirically and reflects the high variance exhibited by speech data and the lower level of variance in silence or non-speech sounds.

In the second step, these models are iteratively used to resegment the original stream into speech, silence and non-speech sound regions respectively and retrain the models. The third step is a simple check of how “similar” the non-speech sound and silence models are. If their similarity is above a threshold, the framework assumes that there are few non-speech sounds, and repeats the procedure assigning all the non-speech data to the 7 component GMM (silence).

Using the speech/non-speech detector, the non-speech windows are detected and discarded. The rest of the stream is assigned to the speakers through clustering. The initialisation has a large number of clusters, which must be much larger than the total number of speakers. Considering that most recordings do not have more than eight participants, the initial number of clusters is usually set to 40. After this initial clustering the score of merging two clusters is computed using the  $\Delta\text{BIC}$  criterion, an idea introduced in earlier work of Wootter et al. [131]. Two clusters parameterised by  $\theta_i$  and  $\theta_j$  can be merged in a single cluster parameterised by  $\theta$  with a merge score of:

$$\text{MergeScore}(\theta_i, \theta_j) = \log p(\mathbf{x}|\theta) - (\log p(\mathbf{x}_i|\theta_i) + \log p(\mathbf{x}_j|\theta_j)) \quad (2.8)$$

where  $\mathbf{x}_i$ ,  $\mathbf{x}_j$  and  $\mathbf{x}$  are the data assigned to the first, second and merged cluster respectively. The most similar clusters are merged and the procedure is repeated until no merging gives a positive score. Each final cluster corresponds to a different single speaker — and the number of clusters corresponds to the number of speakers in the final recording.

## Conclusions

The audio-based speaker diarization can be applied to all the possible speaker diarization contexts. Simple statistical measurements can provide an adequate estimation of speaker change points, and intuitive clustering techniques can provide speaker-homogeneous clusters. This is evidence that speech has well organised statistical properties that can be exploited for speaker diarization.

The results reported by different researchers show low DER, smaller than 15%, but in scenarios where the problem has been simplified greatly [50, 132]. The labelling windows often have coarse discretisation, and collars around the human annotation discard the most challenging parts of the stream. Precise speaker diarization is especially challenging in the following situations: (1) positions where precise speaker change detection is hard

and (2) short utterances, which contain little acoustic information to identify the speaker. These situations are not only challenging, but also very common in natural dialogues and applications like for example HCI and video conferences.

The audio-based methods, however, ignore the video modality of a recording. Information coming from the video stream can be used to improve the speaker diarization accuracy over the whole stream, and to provide precise results for the hard parts of the stream. The next section describes methods that use both the audio and video modality in the simplest way: they detect which region of the video stream is most synchronised to the audio stream and assume that this corresponds to the speaker.

## 2.3 Synchrony based Speaker Diarization

Using the synchrony between audio and video, humans can locate which part of their visual input is likely the source of audio input and decide whether their perception of a speaking person is synchronised to a specific audio stream. In order to enable machines to perform synchrony detection and apply it to speaker diarization, relevant research has split the problem of synchrony detection in three steps: (1) extract features from the audio and video stream (2) measure the synchrony using these features and (3) use this synchrony measurements to detect the speaker.

Relevant research makes three implicit assumptions. First, the features extracted from the audio and video stream are assumed to contain the synchrony-related information. Second, the measure of synchrony is assumed to correspond to how likely it is that the two streams are synchronised. Third, the person appearing most synchronised to the audio stream is assumed to be the speaker.

The current section categorises relevant research in two categories, namely the Mutual Information (MI)-based approaches [59, 66] and the matching algorithm-based approaches. The MI-based approaches extract high-dimensional low-level features from the two signals with minimal processing, e.g., intensity video pixels or audio energy. Then, they implicitly assume that the MI between the audio and video features reflects audiovisual synchrony: the higher the MI, the more synchronised are the original streams. The output of those frameworks varies from visual detection of the audio source [59, 66], to source separation [66] and speaker detection [59, 66].

The matching algorithm-based approaches [12, 73] process the audio and video signals extensively in order to extract low-dimensional high-level features such as the detection of sudden changes in the audio stream, or the acceleration of distinctive visual features. The assumption made is that these features are low-dimensional but contain all the necessary information for synchrony detection. Complex models or matching algorithms can be applied to capture the low-dimensional feature relationships, which are assumed to represent synchrony between the audio and video stream. The output of those frameworks

varies from visual detection of the audio source [12], to source separation [12] and speaker detection [12, 73].

The remainder of this section is organised as follows: section 2.3.1 reviews the approaches which detect synchrony based on MI. Section 2.3.2 reviews the matching algorithm-based approaches. Section 2.3.3 describes how the output of different methods can be used to perform speaker diarization.

### 2.3.1 Mutual Information-based Methods

The earliest effort to measure audiovisual correlation is the work of Hershey and Movellan [59], which is bound to the assumption that high MI between the audio and video features reflects synchrony between the audio and video modalities. Intuitively, MI between variables  $\mathbf{X}$  and  $\mathbf{Y}$  measures the information about  $\mathbf{X}$  that is provided by  $\mathbf{Y}$ . It is denoted as  $MI(\mathbf{X}; \mathbf{Y})$  and it is given by:

$$MI(\mathbf{X}; \mathbf{Y}) = \int_{\mathbf{x}} \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log \left( \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \quad (2.9)$$

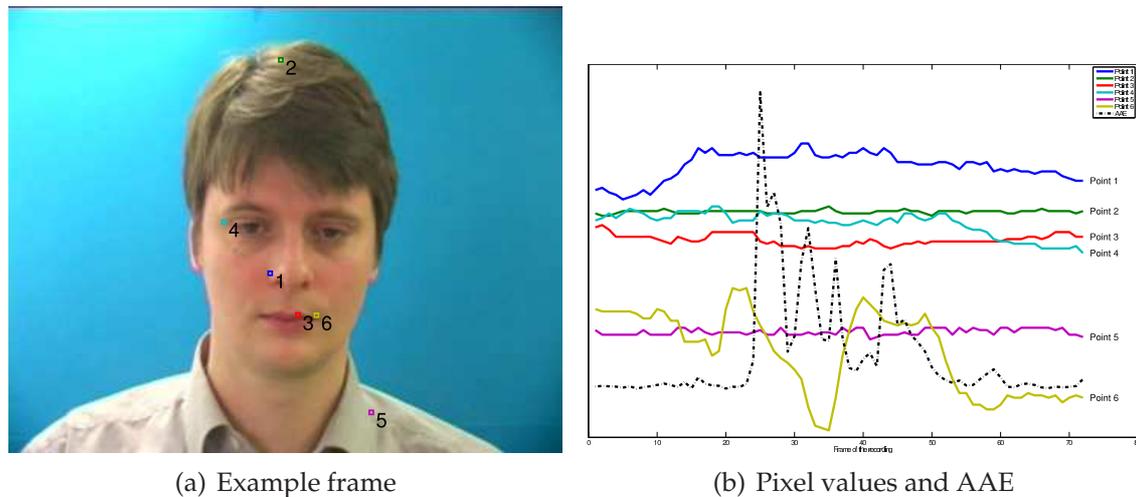
Hershey and Movellan suggest the estimation of the MI between the pixel values and the average acoustic energy of the audio stream. In general, MI can not be computed explicitly in closed form. However, assuming that variables  $\mathbf{X}$  and  $\mathbf{Y}$  are Normally distributed, there exists a closed-form expression of their MI:

$$MI(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \log \left( \frac{|\Sigma_{\mathbf{X}}| |\Sigma_{\mathbf{Y}}|}{|\Sigma_{\mathbf{XY}}|} \right) \quad (2.10)$$

where  $\Sigma_{\mathbf{X}}$  and  $\Sigma_{\mathbf{Y}}$  are the covariances of the distributions of the variables  $\mathbf{X}$  and  $\mathbf{Y}$  respectively and  $\Sigma_{\mathbf{XY}}$  the covariance of their joint distribution.

The assumption of a Gaussian distribution was used in the original paper of Hershey and Movellan [59], and has been very influential: Darrel et al. in [42] based their synchrony detection results on the same method, and Harriet, Giridharan and Calapathy performed multiple subject experiments in [56] and extensive monologue and speaker detection experiments on publicly available data sets in [65], all assuming that the extracted audio and video features have a Gaussian distribution within the window they consider.

It follows that the transformation from audio and video signals to random variables, that is, the feature extraction procedure, becomes critical. The ideal features would have a Gaussian distribution and should be low-dimensional with respect to their sampling rate. This is because only a short temporal window is examined for synchrony, and the samples of that window should give a good estimate of the features' covariance. For example, the audio modality usually provides tens of thousands of samples per second and, therefore, we can have a relatively high-dimensional feature vector and still estimate the data covariance



**Figure 2.3:** On the left, an example frame from a video sequence with 6 pixels selected, coming from the nose, the hair, the eye, the shirt and the lips of the person. On the right the gray-scale value variation for the selected pixels over 72 frames, as well as the average acoustic energy of the audio stream over the same period.

reliably with a few seconds of data. However, the video modality usually provides 25 frames per second making high-dimensional descriptors ill-suited: we cannot reliably estimate the covariance of a high-dimensional variable using a few samples.

In the work of Herhsey and Movellan, the MI was estimated between each independent pixel's intensity variation and the Average Acoustic Energy (AAE) of the corresponding audio stream. The AAE of an audio window is estimated as the sum of the absolute or the squared values of its samples. In order to acquire a measure for a set of pixels, e.g., a window in the frame, the average MI of the pixels of the set is used. In figure 2.3(a) an example frame of a 72-frame sequence of a speaking person is presented. The gray-scale values of 6 different pixels, which are on fixed positions over the 72 frames, as well as the AAE of the corresponding audio stream are plotted in figure 2.3(b). The pixel coming from the edge of the lips (point 6) exhibits the highest variation while the rest of the pixels exhibit little variation. Notice that the corresponding audio stream also exhibits variation at the same time that the pixel coming from the edge of the mouth does. However, a nearby pixel (point 3) does not exhibit a similar behaviour.

Other MI-based approaches use different feature extraction methods. For example, Darell et al. whitened the image in order to avoid illumination-related variations [42]. In the audio modality, Iyengar et al. proposed the use of MFCC instead of the AAE of the audio stream [56], with the expectation that MFCCs contain potentially useful phoneme-specific information. The choice of audiovisual features is limited though, because synchrony detection for a specific point of the stream can only use information from a short temporal window; this prohibits the use of high-dimensional features which would require the estimation of

high-dimensional covariance matrices.

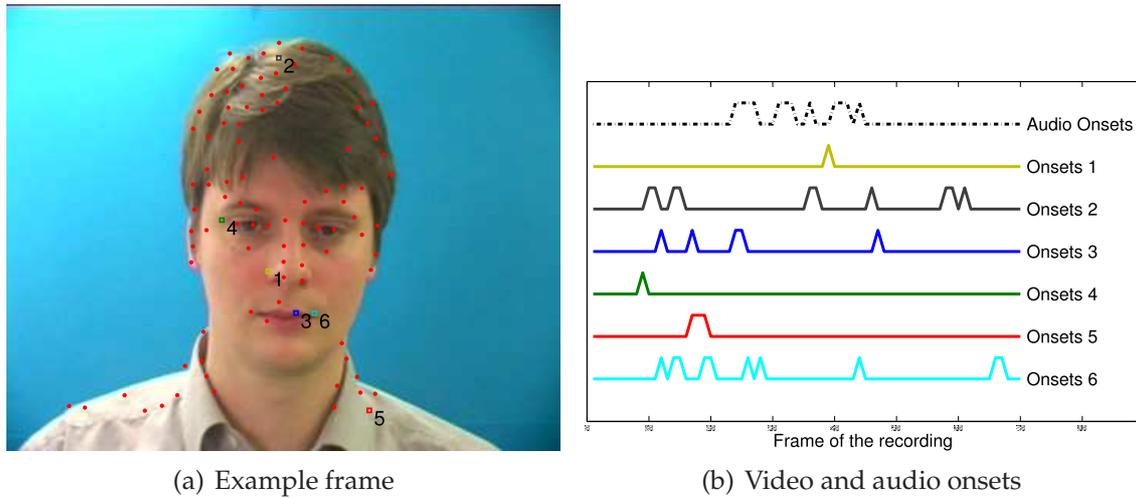
Fisher and Darrel in [66] relax the parameasure assumptions of the original approach about a Gaussian distribution. The proposed algorithm discovers the optimal linear projection of the audio and video descriptors in an one-dimensional subspace where the MI of the two projections is maximised. Once the projection's coefficients are recovered, the audio signal is mapped to the one dimensional space and from there to a set of pixels in the video frame. These pixels are expected to correspond to the source of the audio. This approach, however, is extremely expensive computationally and requires complex heuristics for its initialisation. Therefore, it is not applicable to real-world problems and the extensive experimental research presented by Harriet, Giridharan and Calapathy in [65] focuses explicitly on the application of MI-based correlation measurements based on the Gaussian assumption.

### 2.3.2 Matching algorithm-based approaches

Previous research also explored synchrony detection on high-level features, i.e., features for the extraction of which extensive processing of the input signals is required. In this line, Barzelay and Schechner in [12], who extend the earlier work of Kidron et al. [73], seek correspondence between significant features in the audio and video streams. This a choice motivated by biological neural systems research concluding that cross-modality association is based on salient features [51]. In synchrony detection, the characteristics of significant features are saliency, reliable detection and high correlation in the audio and video modality. In the work of Barzelay and Schechner, the features regarded significant are onsets in the video and audio modality. Onsets in the video and audio modality are points in the stream where each signal exhibits strong temporal variation [12].

In the video modality, the first step is to detect features that can be tracked over multiple frames. In the works mentioned above, Kanade-Lucas-Tomasi (KLT) features are used. KLT features are located by examining the minimum eigenvalue of each two by two gradient matrix, and they are tracked using a Newton-Raphson method of minimising the difference between two consecutive windows. Multi-resolution tracking allows for relatively large displacements between images. The original idea for such tracking dates back to 1981 and the work of Lucas and Kanade [81], and the implementation used by Barzelay and Schechner was further developed in the works of Tomasi and Kanade [123] and Shi and Tomasi [113]. In figure 2.4(a) the 100 best features for the same video sequence as figure 2.3 are shown.

In order to decide when an onset occurs, each feature  $i$  is tracked independently. The magnitude of the feature's acceleration at frame  $t$  is measured, thresholded and temporally pruned. This results in a binary vector  $\mathbf{v}_i$  for each feature  $i$ , where element  $v_i(t)$  is one if feature  $i$  has high acceleration at  $t$  and zero otherwise. In figure 2.4 the onset vectors of six selected features are shown — the selected features correspond to the points whose



**Figure 2.4:** On the left, an example frame with the one hundred best features to track for a 72 frames sequence superimposed. On the right the onset vector extracted from tracking 6 of these features.

gray scale value variation was shown in figure 2.3(b) .

In the audio modality, onset detection is a well-studied problem; see for example the tutorial of Bello et al. [15]. In the work of Barzelay and Schechner, the detected onsets are based on psychoacoustic knowledge as described in the work of Klapuri [74]. In short, the initial audio signal is divided into 21 non overlapping bands which are full-wave rectified and convolved using a half-Hanning window. Onset detection is performed in each band independently, by locating the peaks in the first derivative of the logarithm of the amplitude envelope. In the final step, the algorithm computes the sum of the onset intensities coming from detections in all the banks. In parallel to the processing of the video modality, the total intensity for each candidate onset is thresholded to provide the onset locations in a vector  $\mathbf{a}(t)$ . The detected onsets for the AAE plotted in figure 2.3(b) are shown in figure 2.4(b).

Barzelay and Schechner perform synchrony detection in the onset space. The matching criterion is defined as:

$$L(i) = 2 [\mathbf{a}^T \mathbf{v}_i] - \mathbf{1} \mathbf{v}_i \quad (2.11)$$

where  $\mathbf{1}$  is a row vector with all elements equal to one. The feature point with the highest synchrony is selected as the source of the corresponding audio stream. Barzelay and Schechner propose the same type of onsets and matching algorithm for a variety of contexts, ranging from the lip motion of a speaker and the corresponding speech to the motion of the violin fiddle and the corresponding melody. However, in their work they report results for only one sequence containing speech [12].

### 2.3.3 From synchrony detection to speaker diarization

The synchrony detection methods described above, propose an approximation to the probability  $p(\text{synchrony}|\text{data})$ , i.e., how probable it is that the observed data reflect synchrony. Clearly, a very good way to demonstrate the applicability of synchrony detection algorithms is to use them for speaker diarization — detect that the animation of a speaker is synchronised to the corresponding speech. Synchrony detection methods perform speaker diarization through the implicit assumption that, at each point of the stream, the person appearing most synchronised to the audio stream is the speaker. In practice this is usually demonstrated on a video showing two persons who speak in turns. The stream is divided into windows of predefined length, and the person appearing most synchronised is selected as the speaker for each window.

This was the choice of experiment in the works of Hershey and Movellan in [59], Barzelay and Schechner in [12], Darrel et al. in [42], and extended experiments of Hariett et al. in [65]. The works of Hershey and Movellan and Barzelay and Schechner present qualitative results of speaker detection on a single two-person recording. Specifically, Hershey and Movellan plot the ground truth and the ratio of MI between the two speakers. Barzelay and Schechner demonstrate their results in a new video, presenting each speaker separately with the corresponding assigned audio stream. Although no quantitative accuracy is reported, the qualitative results seem nearly perfect.

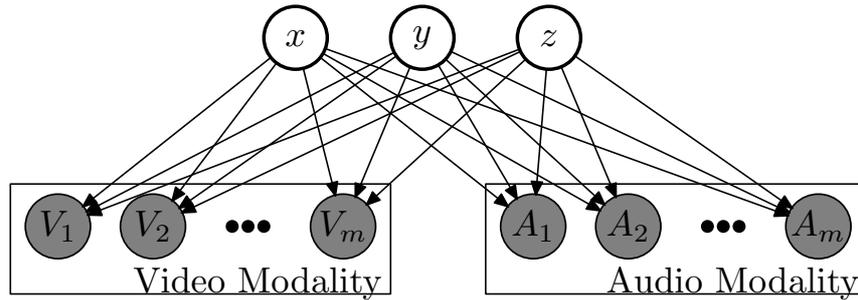
Hariett et al. perform extensive evaluation on multiperson real-world recordings coming from TRECVID contest videos or teleconferencing [65]. They report accuracy between 50% and 82% depending on the context and the number of speakers. A more challenging problem was presented in the work of Fisher and Darrel in [66], where eight individuals were recorded speaking, and all the possible video-to-audio combinations were evaluated. Each video was matched to the audio, with which they maximise their MI measurement.

## Conclusions

Relevant research on synchrony detection tackled synchrony in a variety of contexts, e.g., musical instruments or speech. The development of methods which are focused on audiovisual synchrony solely in speech remains a challenging field. Chapter 4 presents a Deep Belief Network that can capture such synchrony.

Moreover, previous synchrony detection methods have demonstrated qualitative results in audiovisual recordings containing speaking persons. However, there has been little quantitative evaluation on how well these methods perform on speaker detection, and how they can be adapted to such a task. Chapter 5 presents a neural network to perform speaker detection, that is based on the Deep Belief Network of chapter 4.

Finally, there has been no effort to combine synchrony detection algorithms with other modalities for speaker diarization. Chapter 3 presents a Dynamic Bayesian Network that



**Figure 2.5:** A graphical model representing probabilistic audiovisual speaker localisation. The coordinates of the speaker in the 3-D world affect the observations in each one of the  $m$  cameras of the video modality, as well as the time differences in the audio input of  $m$  microphones.

can incorporate synchrony detection methods to improve speaker diarization results and facilitate audiovisual fusion.

## 2.4 Localisation-Based Speaker Diarization

When humans perform speaker diarization they use both their sight and hearing. It therefore seems natural to employ systems that combine information from their audio and video input. The development of such systems is inherent to single modality analysis and the difficulties arising in low-level audiovisual fusion. Relevant research on audiovisual speaker diarization proposes probabilistic models to deal with this uncertainty and perform speaker diarization using information from both modalities, in the form of speaker localisation.

The general probabilistic model of speaker localisation is sketched in figure 2.5. The speakers' location affects both the audio signal measured by a set of microphones (and more specifically the difference between the phase and amplitude of their measurements), as well as how they are seen by the cameras. Therefore, when multiple persons are accurately located using the video modality, the localisation of the source of the audio stream corresponds to the most probable speaker. On the one hand, more cameras and microphones provide additional sources of information and potentially better results. On the other hand, multiple input streams are not present in most of the existing recordings and complicate the processing procedure. Most of the proposed approaches, therefore, simplify the model presented in figure 2.5.

### 2.4.1 Practical Implementation

The first framework proposed to perform audiovisual speaker localisation was the work of Beal et al. [13]. In this work, a speaker is walking in front of a camera and the objective

is to locate his position in the horizontal axis. The time difference in the input of two microphones is estimated and consequently the orientation of the audio source. In the video modality the speaker's position is inferred on the basis of a simple colour-based appearance model. The framework proposed by Beal et al. requires no labelled training data. Instead, the Expectation Maximization (EM) algorithm is used, in order to acquire the model parameters and the position of the speaker directly from the test data. The work of Beal et al. presented only qualitative results in the form of a demonstration video. In this video the multimodal approach appears more stable than the audio-only tracking. Moreover, the multimodal tracking approach does not lose track of the position of the speaker because of occlusions, in contrast to the video-based tracking.

The same idea was used by Cutler et al. in [39] and Chen et al. [32], where information from multiple cameras and a set of microphones is used to perform speaker localisation on-line. In short, a particle filter maintains different hypotheses about the location of the speaker, and the framework updates all the particles as information from the audio and video modality becomes available. The results of a particle filter are of high quality when (1) the assumptions about the temporal behavior of the data hold and (2) when good novel hypotheses are generated, in case the proposal distribution is not close to the posterior distribution of the next step.

Cutler et al. used task-specific hardware to capture the necessary video information [39]. An inexpensive omnidirectional camera is set in the middle of a table, around which a meeting takes place. This camera provides the locations of different detected participants, which are the possible speakers at each point in time. The differences between different microphones provide a noisy estimation of the location of the audio source and the person detected closest to the current estimation is selected as the speaker. Cutler et al. use their system to enhance human-to-human communication in a video conference setting and report qualitative results in the form of feedback by the participants. These results are in favour of the proposed system and indicate the potential applicability of speaker diarization to assist video conferences.

Chen et al. processed the input of off-the-shelf camera equipment using a contour tracker and a color tracker [32]. The particle filter incorporated the output of those trackers rather than the visual input signal directly. In the audio modality, the time difference between the available microphones is used to detect the location of the speaker. The final framework of Chen et al. had another level of processing, named "fuser", which corrected small errors made by the particles maintained at each point of time. They report "robust" speaker diarization results, but give no quantitative measurement.

Finally, the works of Checka et al. [29] and Gatica-Perez et al. [54] focused on off-line people tracking and speaker diarization in meetings respectively. In these works, a set of calibrated cameras track and detect the visible persons. In the work of Checka et al. the background is subtracted to locate the position of the speakers while all the sound measurements are analysed to locate audio sources — such as speech or steps. The probabilistic model infers the most probable set of correspondences between audio sources

and the detected persons. The audio and video signal of each individual are used to estimate their accurate position. In a single experimental video containing two speakers, the first speaker is correctly identified in 90% of the cases and the second speaker in 84% of the cases, while non-speech parts are correctly identified with 65% accuracy.

In the work of Gatica-Perez et al., the heads of the meeting participants are tracked. This is very useful information because a lot of visual ASR information can be recovered from the speakers face, but makes the system more vulnerable to occlusions. The audio information provides the 2-D location of the speaker and it is derived from a microphone array. A complex model tracks both the individual participants and the multi-speaker interactions. Since this creates many complex dependencies, inference about the active speaker is performed using Markov Chain Monte Carlo (MCMC) sampling. A microphone array uses the time differences between the audio inputs to locate the most probable source of audio at each point of the recording and the person identified closer to the location is considered the active speaker. This method improves over simple tracking algorithms and achieves more than 80% speaker diarization accuracy for all the speakers.

## Conclusions

Relevant research in audiovisual speaker diarization approaches treats the problem as a multimodal tracking application. On the one hand, this produces high-accuracy speaker diarization, which is based on a solid mathematical framework. On the other hand, multiple cameras and multiple microphones are not available in most recordings, and in most of the existing digital libraries there is little or no information about the location of the recording equipment. Moreover, the actual synchrony between the audio and video modalities is barely used, the audio and video stream are treated as independent sources of information.

In chapter 3 an alternative Dynamic Bayesian Network is proposed to perform multimodal speaker diarization. The proposed model explicitly models the voices and appearances of different persons, which does not require multiple microphones or knowledge about the location of the equipment. The audiovisual fusion is performed under a probabilistic framework which exploits the synchrony between audio and video stream, and acquires the model parameters online from the test data.

## 2.5 Conclusions

Speaker diarization is treated by relevant research as (1) an audio problem, (2) an audiovisual synchrony problem or (3) an audiovisual localisation problem. Audio-based speaker diarization has examined different acoustic features and modelling choices. These choices lead to systems that can be very robust in high-quality recordings where different speakers

can be easily distinguished. Audio-based speaker diarization remains suboptimal, however, since the information of the video modality is ignored.

Synchrony-based audiovisual speaker diarization is based on assumptions about the parametric distribution of the audiovisual features, or the ability of simple onset features to capture the necessary information. However, none of these methods is focused on audiovisual synchrony in speech but, in contrast, treat it the same way as for example synchrony between the motion and music produced by a violin player. Moreover, these methods assume that the output of synchrony detection can be directly mapped to speaker detection; an assumption that does not hold in practice.

Finally, localisation-based audiovisual speaker diarization treats the problem as a tracking task, where the speaker is simultaneously tracked in the audio and video modality. Such approaches often require a set of calibrated microphones and cameras which are not available in most of the recordings. Furthermore, no person-specific information is incorporated (e.g., the appearance or voice of each person).

Previous research has made numerous advancements towards human-like speaker diarization, but the task is far from solved. This thesis addresses three open issues in speaker diarization. Chapter 3 proposes a probabilistic model that can fuse information coming from the audio, video and audiovisual space and perform high-accuracy speaker diarization. Chapter 4 presents a Deep Belief Network which directly models synchrony detection in speech. Chapter 5 describes how to move from synchrony detection to speaker detection, and how to incorporate this output to speaker diarization.