Multimodal Speaker Diarization

Athanasios Noulas, Gwenn Englebienne, and Ben J.A. Kröse, *Member*, *IEEE*

Abstract—We present a novel probabilistic framework that fuses information coming from the audio and video modality to perform speaker diarization. The proposed framework is a Dynamic Bayesian Network (DBN) that is an extension of a factorial Hidden Markov Model (fHMM) and models the people appearing in an audiovisual recording as multimodal entities that generate observations in the audio stream, the video stream, and the joint audiovisual space. The framework is very robust to different contexts, makes no assumptions about the location of the recording equipment, and does not require labeled training data as it acquires the model parameters using the Expectation Maximization (EM) algorithm. We apply the proposed model to two meeting videos and a news broadcast video, all of which come from publicly available data sets. The results acquired in speaker diarization are in favor of the proposed multimodal framework, which outperforms the single modality analysis results and improves over the state-of-the-art audio-based speaker diarization.

Index Terms—Speaker diarization, dynamic Bayesian networks, audiovisual fusion.

1 Introduction

Speaker diarization corresponds to the task of segmenting a digital recording in speaker homogeneous parts and assigning each part to the corresponding speaker [1]. The output of speaker diarization is useful for Automatic Speech Recognition (ASR) and automatic transcription. In ASR, a generic model of speech is adapted to each speaker. Speaker diarization can provide speaker-specific data for this adaptation in a multispeaker setting. In automatic transcription, the output of speaker diarization can organize the transcript in terms of the speaker's identity. Such a transcript is more readable by humans and is more useful for machines [2].

In speaker diarization, all the available information can be used to identify the speaker at each part of the recording, e.g., models of voices or silence, information from the rest of the recording, the location of the recording equipment, temporal information, and so on. Consequently, speaker diarization involves elements of signal processing, computer vision, and machine learning.

Previous research, which is described in detail in Section 2, focused on subaspects of this problem. A first line of research performs speaker diarization using only the audio stream [3], [4], [5], [6], [7], [8], [9], [10], [11]. In this case, the stream is first segmented at the speaker change positions. Then, assuming a single speaker per segment, the segments are clustered until each cluster corresponds to a single speaker. Such approaches are robust to different contexts, but suboptimal since they do not use the available video information. A second line of research performs speaker diarization through synchrony detection, i.e., they detect the image region in the video frames which is most synchronized to the audio stream and assume it corresponds to the speaker [12], [13], [14], [15], [16], [17], [18]. This is an intuitive assumption, but it often does not hold in practice, e.g., in a recording containing silence, the most synchronized part of the video modality does not correspond to the speaker. Finally, there are approaches which treat speaker diarization as an audio-visual tracking task [19], [20], [21], [22], [23]. The source of the audio is tracked through the difference in phase and amplitude between measurements of different microphones. Potential speakers are tracked in the video modality, and the one closer to the source of the audio is detected as the speaker. Such an approach is based on solid principles, but it is not applicable to the vast majority of available recordings, for which no microphone arrays are used and the speakers are often not visible.

This paper focuses on speaker diarization in audiovisual recordings containing a single audio track and one or more synchronized video streams. This choice of input modalities makes the proposed framework applicable to most of the digital recordings existing today, from web-camera videos to movies and smart meeting room sessions.

We propose a Dynamic Bayesian Network (DBN) that can incorporate information coming from the audio modality, the video modality, and the joint audiovisual space as well as the dynamics appearing in the temporal dimension of an audiovisual stream. The proposed model makes no assumptions about the location of the audiovisual recording equipment, does not assume a single speaker per segment, and acquires the person-specific parameters online, without need for training data.

The remainder of this paper is organized as follows: Section 2 describes the related work which is most relevant to the proposed framework. Section 3 presents the probabilistic formulation of the task of speaker diarization. Section 4 describes how the people of a recording are represented as processes which produce cues in the audio, video, and joint audiovisual space. Section 5 explains how to deal with data association problems arising under the proposed formulation, while Section 6 presents inference

• *The authors are with the University of Amsterdam, Amsterdam, The Netherlands. E-mail: noulas@gmail.com, G.Englebienne@uva.nl, b.j.a.krose@science.uva.nl.*

Manuscript received 6 Sept. 2010; revised 13 Jan. 2011; accepted 10 Feb. 2011; published online 1 Mar. 2011.
Recommended for acceptance by N. Lawrence.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-09-0687.
Digital Object Identifier no. 10.1109/TPAMI.2011.47.

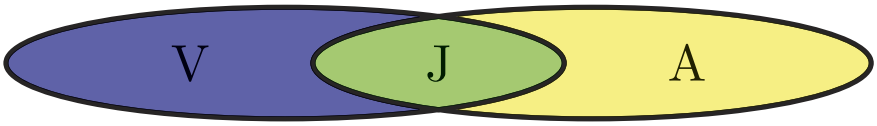


Fig. 1. The different inputs used for speaker diarization. A stands for the audio modality, V for the video modality, and J for the joint audiovisual space.

and learning under the current model. Finally, Section 8 describes the experiments carried out for this work.

2 Related Work

The relevant research can be categorized into three categories, namely, approaches using the audio input (space A), approaches using synchrony (space J), and approaches performing audiovisual tracking (spaces A and V). The possible input modalities are illustrated in Fig. 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, e.g., the location of a potential speaker. The J space corresponds to information coming from the joint audiovisual observations, e.g., whether the motion of a person's lips is in agreement with the phonemes of the audio space.

Using the audio modality alone is the most common choice in the literature, see, for example [3], [4], [5], [6], [7], [8], [9], [10], and led the National Institute of Standards and Technology (NIST) to organize the Rich Transcription (RT) evaluation in which speaker diarization is an independent task. In the RT benchmark of 2007 [2], the method of Wooters and Huijbregts performed best [11] and is still considered the state of the art. It will be used as a baseline for the results of this paper.

Speaker diarization systems based on *synchrony* perform synchrony detection in the J space, i.e., they locate the part of the video modality that appears most synchronized to the audio modality. Assuming that this part of the video modality corresponds to the speaker, the output of synchrony detection can be directly mapped to speaker diarization. Previous approaches on synchrony detection can be divided into two categories. The first category processes the audio and video signals extensively in order to extract features, such as the detection of sudden changes in the audio stream or the acceleration of distinctive visual features. A matching algorithm is used on these changes to locate the parts of the video stream that appear most correlated to the audio stream [12], [13]. The second category involves the estimation of the Mutual Information (MI) between the audio and video signals [14]. The parts of the video stream that are most informative of the audio stream are then selected. Synchrony detection based on MI has been very influential [15], [16], [17], [18]; it was extensively evaluated in recordings containing speech and it will be the choice of this work.

Speaker diarization based on localization performs independent tracking in the audio and video modality. In the audio modality, the source of the audio is located using the difference in the phase and amplitude between measurements of different microphones. In the video modality, the different people are detected and the person closest to the source of the audio stream is located as the speaker. There are multiple speaker diarization frameworks proposed in the literature in which the two modalities are treated independently for tracking [19], [20], [21], [22]. In contrast, audiovisual tracking which models both modalities jointly is rare [23].

3 Model Formulation

This paper presents a DBN which manages to capture patterns appearing not only in a single modality, but also across multiple modalities, as well as in the temporal dimension of the multimodal streams. The proposed DBN is constructed as follows:

1. The temporal patterns are treated by a factorized transition model.
2. The state of different speakers is represented by the hidden nodes.
3. The model parameters capture the way persons affect the audiovisual stream.
4. Performing speaker diarization maps to infer the state of the hidden variables.

3.1 Transition Model and Hidden System State

The proposed DBN has a time slice duration equal to the frame duration of the audiovisual recording and takes into consideration the system state in the previous time slice under a factorized transition model. Factorized transition models were first introduced in the factorial Hidden Markov Model (fHMM) [24], which is a constrained version of the Hidden Markov Model (HMM) [25].

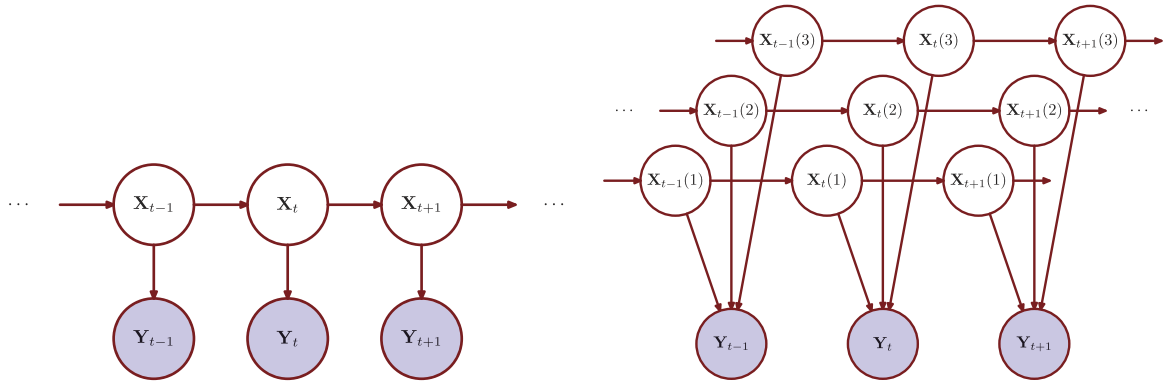
In a first order HMM, as shown in the graphical model of Fig. 2a, the hidden system state at time t, $\mathbf{X}_t$, is discrete[1] and only depends on the previous system state $\mathbf{X}_{t-1}$. In an fHMM, the hidden state is divided into subsets of hidden variables, as shown in Fig. 2b. Consequently, the transition probability of the system state *factorizes* into a product of terms, each one of which depends on a subset of the variables of $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$. In probabilistic terms, this corresponds to

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \prod_i p(\mathbf{X}_t(i)|\mathbf{X}_{t-1}(i)), \quad (1)$$

where $\mathbf{X}_t = (\mathbf{X}_t(1), \mathbf{X}_t(2) \dots \mathbf{X}_t(n))$. This factorization leads to a drastic decrease in the number of free parameters for the transition probabilities.

In the proposed system, the hidden system state $\mathbf{X}_t$ represents the identity of the speaker(s) and the visible person(s) at time t. In case of N people, each realization $\mathbf{x}_t$ is a binary vector of length $2N$, where elements $[1 \dots N]$ are 1

1. In this paper, capital letters denote random variables and lowercase letters denote instantiations. Bold symbols denote vectors.



(a) The Hidden Markov Model (b) The factorial Hidden Markov Model

Fig. 2. The graphical models of a Hidden Markov Model and a factorial Hidden Markov Model.

if the corresponding person is visible and 0 otherwise, and elements $[N+1 \dots 2N]$ are 1 if the corresponding person speaks and 0 otherwise. The first N elements are denoted with $\mathbf{X}_t^V$ and correspond to the video part. The second N elements are denoted with $\mathbf{X}_t^A$ and correspond to the audio part. Note here that we assume a known number of speakers—we compare models with different number of speakers to automatically select the correct number in Section 8.2. The transition probability is factorized in terms of these variables as

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \prod_n p\big(X_t^A(n)|X_{t-1}^A(n)\big)p\big(X_t^V(n)|X_{t-1}^V(n)\big), \quad (2)$$

which implies that the transition probability for the state of each person is independent of the state of the other persons. For example, the fact that a person becomes visible on a specific frame is independent of whether he starts to speak.

The binary representation of the hidden system state creates a total of 2^{2N} possible system states. The transition matrix of a DBN which explicitly defines $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ would contain 2^{4N} parameters. Factorizing the system state in independent person states decreases the number of free parameters for the transition matrix. In particular, $X_t^A(n)$ and $X_t^V(n)$ are binary and we need two parameters for each factor.[2] Thus, the factorized transition matrix is defined by just $4N$ parameters in total.

3.2 Observations and Observation Model

The visible nodes at time t, denoted with $\mathbf{Y}_t$, depend on the system state of time t. In the proposed model, the observation $\mathbf{Y}_t = (\mathbf{A}_t, \mathbf{V}_t, N_t^f, \mathbf{J}_t)$ represents the features extracted from the multiple modalities, namely, the audio stream ($\mathbf{A}_t$), the video stream ($\mathbf{V}_t, N_t^f$), and the joint audiovisual space ($\mathbf{J}_t$) at the corresponding time t.

In an HMM, the observation model consists of the conditional probability $p(\mathbf{y}_t|\mathbf{x}_t)$. In an fHMM, this distribution cannot be factorized in a general way. In the proposed DBN, however, $p(\mathbf{y}_t|\mathbf{x}_t)$ is factorized into one observation model per person, called *person model*. The size and type of a realization of the observations $\mathbf{y}_t$ and the type of the person models are dependent on our choice of features, but any kind of feature can be incorporated under the proposed framework. The feature choices of this work can be found in Section 4, while the specific factorization for the observation model is presented in Section 6.

2. In order to parameterize $p(X_t(n)|X_{t-1}(n))$ and $p(X_t(n)|\neg X_{t-1}(n))$.

3.3 Parameterization

The proposed model is defined by the priors π for each state, the transition matrix $\mathcal{A}$, and the observation model. The parameter $\pi(\mathbf{x})$ represents the probability of the system being in state $\mathbf{X}_1 = \mathbf{x}$ at time step $t = 1$. The transition matrix is factorized using person-specific factors $\mathcal{A}_{ij}^{nA} = p(x_t^A(n) = j|x_{t-1}^A(n) = i)$ and $\mathcal{A}_{ij}^{nV} = p(x_t^V(n) = j|x_{t-1}^V(n) = i)$. Element $\mathcal{A}_{\mathbf{ij}}$ denotes the probability of transition from state $\mathbf{X}_t = \mathbf{i}$ to $\mathbf{X}_{t+1} = \mathbf{j}$.

The graphical model representation of the proposed framework is depicted in Fig. 3. In this model, the individual person models are independent and their transition probabilities factorize. However, the observation at time t depends on all the person states at that point in time, and therefore the hidden states of t are not conditionally independent of each other given the observations $\mathbf{y}_{1:T}$. Note here that the simple model in Fig. 2b is not directly applicable to the extracted features. This proposed network structure, depicted in Fig. 3, is necessary for the solid probabilistic treatment of multiple persons—explained in Section 5.

3.4 Inference

Given a new audiovisual stream of T time slices, the problem, in probabilistic terms, translates to estimating the state sequence $\mathbf{x}_{1:T}$ that best "explains" the observation sequence $\mathbf{y}_{1:T}$ extracted from that stream [25]. This will return the identity of the speaker and the visible people at each point of the stream. Inference is described in Section 6.

4 Person Models

People generate features in the video and audio streams, as well as in the joint audiovisual space. Let the parameters of the person model of the nth participant be $\boldsymbol{\theta}_n$, consisting of three parts: $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_n^V, \boldsymbol{\theta}_n^A, \boldsymbol{\theta}_n^J)$. $\boldsymbol{\theta}_n^V$ denotes the **V**ideo modality part, $\boldsymbol{\theta}_n^A$ the **A**udio modality part, and $\boldsymbol{\theta}_n^J$ the part concerning the **J**oint audiovisual space. For example, the

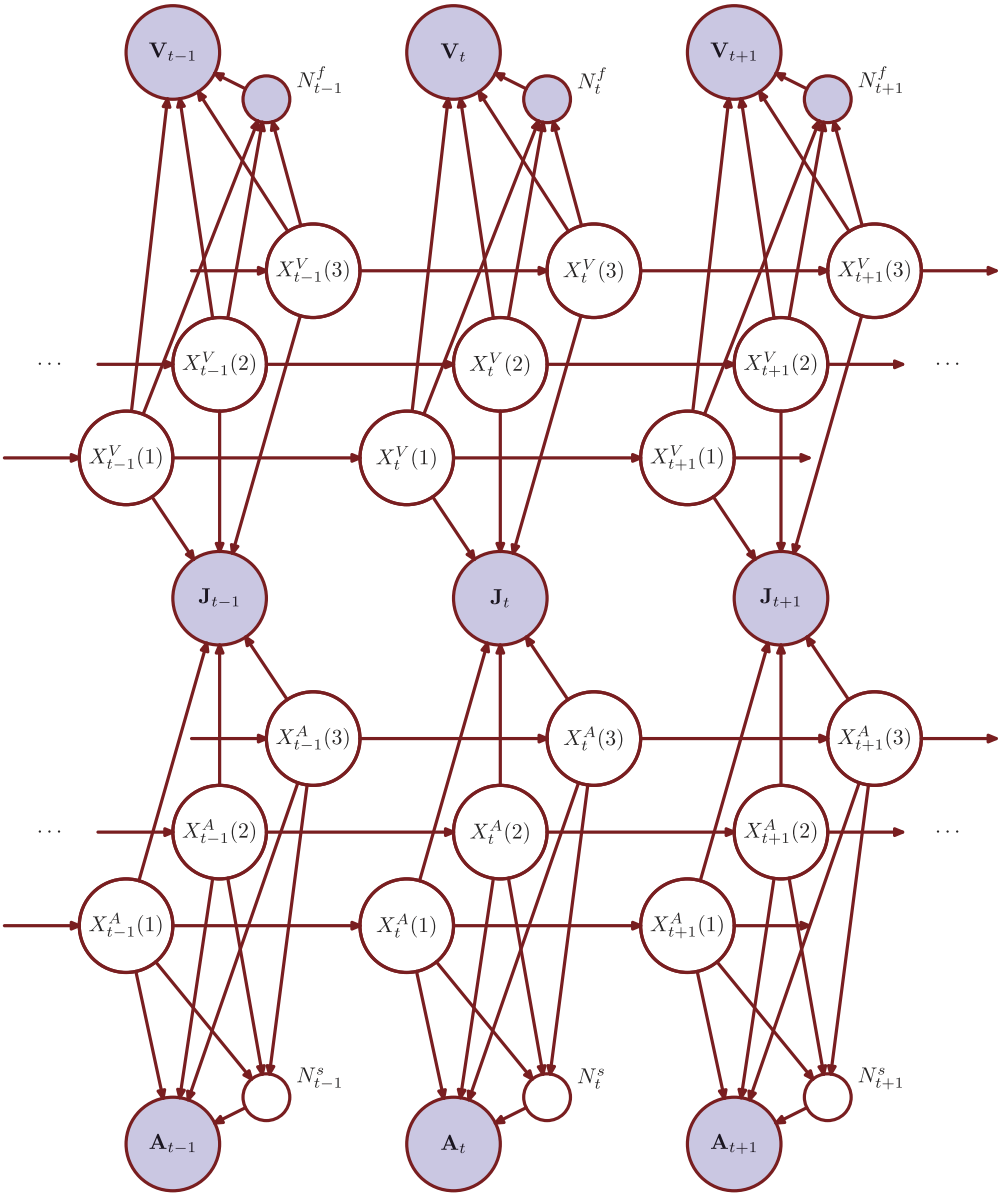


Fig. 3. The dynamic Bayesian network used for audiovisual fusion for speaker diarization. Shaded nodes represent observed variables, while unshaded nodes represent latent variables. The person states in both modalities are dependent in the observation model and independent during transition. The nodes N^f and N^s denote the number of detected people and the number of speakers, respectively, and allow us to model the video and audio modality in a generative fashion. Note that we estimate the distribution over N^s during inference, and therefore the node variable is latent, while the distribution over N^f is estimated in a preprocessing step, and therefore it is denoted as an observable variable.

probability that observation $\mathbf{v}_t$ was generated in the video modality, given that person 2 was visible, is

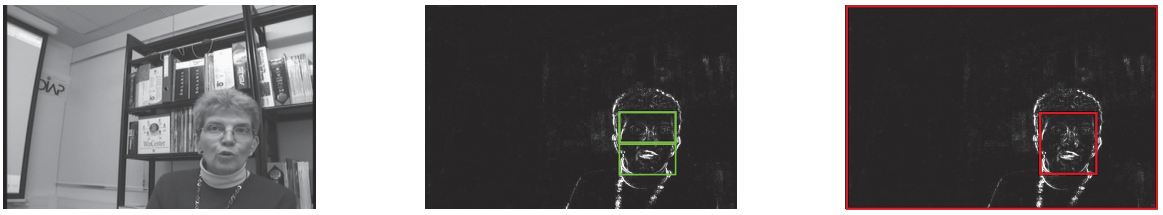
$$p\big(\mathbf{v}_t|\mathbf{x}^V_t(2)=1\big)=p\big(\mathbf{v}_t;\boldsymbol{\theta}^V_2\big). \tag{3}$$

These parameters represent the realization of a generative distribution in the feature space of each modality and correspond to the probability that an observation $\mathbf{y}_t$ is generated by the corresponding person. Thus, the family of the distribution depends on the type of the extracted features and is fixed beforehand, while the parameters of the distribution are learned from the data.

4.1 Video Space

In the *video modality*, the regions of interest are faces, which are detected using the Viola-Jones face detector [26]. The face descriptors are extracted using the *Bag of Keypoints* method [27] which, in short, works as follows: Scale Invariant Feature Transforms (SIFTs) appearing in these regions are extracted, and vector quantization in the SIFT space is performed. The chosen number of clusters, often described as "visual words" [27], is set manually.[3] In this work, 100 visual words were used. Each face region, based on the output of the region-of-interest detection, will return a different number of SIFT descriptors. Each descriptor is assigned to the closest cluster, and the final observation extracted from the video modality of the stream (denoted with $\mathbf{V}_t$) is a binary vector of length 100, with each element denoting the existence (value 1) or absence (value 0) of the

3. Unless extreme values are set, there is no significant change in the accuracy of our framework.



(a) (b) (c)

Fig. 4. (a) The frame of the clip and (b)-(c) the Mutual Information Image. Green rectangles denote the upper and lower halves of the detected face (b). Red rectangles denote the face and the whole frame regions (c).

corresponding visual word in the face region—there is one such vector per detected face in each frame.

The video modality part of the person model is modeled by a set of Bernoulli distributions, representing the probability that a specific visual word is present in the face region given the person's identity. That is, $\boldsymbol{\theta}_n^V = (b_1^n, b_2^n \dots b_{100}^n)^\top$. We assume independence among the appearance of visual words so that:

$$p(\mathbf{v}_t; \boldsymbol{\theta}_n^V) = \prod_i p(v_t(i); b_i^n) = \prod_i \left((b_i^n)^{v_t(i)} (1 - b_i^n)^{1 - v_t(i)} \right). \quad (4)$$

4.2 Audio Space

In the *audio modality*, the stream is divided in 16-ms windows with a 6-ms overlap. The time slice of the model has frame duration (40 ms) and therefore four audio descriptors are extracted, denoted as $\mathbf{a}_t(m)$ with $m \in \{1 \dots 4\}$, in each audio observation $\mathbf{a}_t$. These descriptors contain the first 13 Mel Frequency Cepstral Coefficients (MFCCs) from the audio stream along with their first and second order differences. The audio part of a person model ($\boldsymbol{\theta}_n^A$) is a 15-component Gaussian Mixture Model (GMM), with means $\boldsymbol{\mu}_i$ and full covariance matrices $\boldsymbol{\Sigma}_i$. The probability of observing an audio feature vector generated by a specific person maps to the corresponding probability density function at that point, $p(\mathbf{a}_t(m); \boldsymbol{\theta}_n^A) = \sum_c \pi_c \mathcal{N}(\mathbf{a}_t(m); \boldsymbol{\mu}_c^n, \boldsymbol{\Sigma}_c^n)$. We assume that consecutive windows are conditionally independent of each other given the person's audio model, and get

$$p(\mathbf{a}_t; \boldsymbol{\theta}_n^A) = \prod_m p(\mathbf{a}_t(m); \boldsymbol{\theta}_n^A) = \prod_m \sum_c \pi_c \, \mathcal{N}(\mathbf{a}_t(m); \boldsymbol{\mu}_c^n, \boldsymbol{\Sigma}_c^n), \quad (5)$$

where $\mathcal{N}(\mathbf{a}_t(m); \boldsymbol{\mu}_c^n, \boldsymbol{\Sigma}_c^n)$ is the evaluation of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_c^n$ and covariance matrix $\boldsymbol{\Sigma}_c^n$ at $\mathbf{a}_t(m)$.

4.3 Audiovisual Space

Finally, the correlations of the *joint audiovisual* space are modeled through the estimate of the MI between the two streams as first introduced in the work of Hershey and Movellan [14]. The MI between two variables, $\mathbf{A}$ and $\mathbf{V}$, measures the information of variable $\mathbf{A}$ that is shared with variable $\mathbf{V}$ and it is defined as

$$MI(\mathbf{A}, \mathbf{V}) = \int_{\mathbf{a} \in \mathbf{A}} \int_{\mathbf{v} \in \mathbf{V}} p(\mathbf{a}, \mathbf{v}) \log \frac{p(\mathbf{a}, \mathbf{v})}{p(\mathbf{a}) p(\mathbf{v})} \mathrm{d}\mathbf{a} \mathrm{d}\mathbf{v}. \quad (6)$$

Consider now a set of audio and video (A and V) samples over a number of time slices. Assuming that the samples of each modality come from a multivariate Gaussian distribution, with variances $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_V$, respectively, and joint variance $\boldsymbol{\Sigma}_{AV}$, the MI estimate becomes [14]

$$MI(\mathbf{A}, \mathbf{V}) = -\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_A||\boldsymbol{\Sigma}_V|}{|\boldsymbol{\Sigma}_{AV}|}. \quad (7)$$

In the audio, our samples are measures of the average acoustic energy of the stream and in the video a single pixel's value variation in gray scale. We estimate the variances using 16 samples around the middle frame, which was also the choice used in [14]. At the resolution of our data, a face contains more than 1.000 pixels while the whole frame contains more than 200.000. They thus produce a very high-dimensional MI descriptor which is called the Mutual Information Image (MII)—see Fig. 4. Since this descriptor is extremely high-dimensional, it is commonly processed further, estimating averages or optimal sets of pixels [14], [28]. In our approach, a *two-dimensional binary vector* is extracted as follows:

- The value of the *first feature* depends on the output of the comparison between the average MI of the pixels in the upper half of the face region to that of the pixels in the lower half of the face region—see Fig. 4b. The intuition is that in the case of random movements or head nods, both the upper and lower parts of the face appear synchronized to the stream; on the contrary, in the case of speech, the moving lower half appears more synchronized than the still upper half.
- The *second feature* reflects the output of the comparison between the average MI of the pixels in the detected face region to the average MI of the pixels in the whole frame—see Fig. 4c. The intuition behind this feature is that it detects when the face is "meaningfully" synchronized with the audio, i.e., when it is more synchronized than the random background.

Both comparisons result in a binary output, and the feature vector $\mathbf{J}_t$ associated to the detected face is

$$\mathbf{J}_t = \begin{bmatrix} \texttt{eval}(MI(\text{upperface}) > MI(\text{lowerface})) \\ \texttt{eval}(MI(\text{face region}) > MI(\text{whole frame})) \end{bmatrix}, \quad (8)$$

where the function `eval` evaluates the comparison of the average MI of the pixels of two specified regions into a binary value.

The audiovisual part of a person model is therefore composed of two generative sets, each consisting of two Bernoulli distributions. The first set models the way the values of $\mathbf{J}_t$ are generated when the person is visible and

speaking ($\boldsymbol{\theta}^J_{n1}$), while the second set models how they are generated when the person is visible but silent ($\boldsymbol{\theta}^J_{n0}$). Similarly to (4),

$$p(\mathbf{j}_t;\boldsymbol{\theta}^J_n)=\prod_i p\big(j_t(i);b^n_i\big)=\prod_i\big((b^n_i)^{j_t(i)}\big(1-b^n_i\big)^{1-j_t(i)}\big), \quad (9)$$

where if $x^A_t(n)=0$ then $\boldsymbol{\theta}^J_{n0}$ is used, while if $x^A_t(n)=1$ then $\boldsymbol{\theta}^J_{n1}$ is used.

4.4 Context Modeling

The audiovisual stream is not affected only by the people but also by the context of the recording. For example, when nobody speaks, the audio stream corresponds to the nonspeech environmental sounds. These environmental sounds are modeled indirectly through a distribution over the number of speakers—described in Section 5. Finally, sometimes the face detection falsely detects faces in the background. The distribution of the video features in such a window is modeled by averaging over all the face models. In effect, this enables us to detect windows that are not a face of any of the people.[4]

5 Probabilistic Treatment of Multiple People

The person models described in Section 4 are used to evaluate the probability that an observation was generated by a specific person. Using these models, it is straightforward to compare multiple person models in order to select the one that most probably generated an observation. However, when dealing with multiperson recordings, it is unfeasible to directly compare all possible system states.

For example, consider a stream with two people with corresponding person models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The system state space is a four-dimensional binary vector, where the first element indicates if the first person is visible, the second element indicates if the second person is visible, the third element indicates if the first person is speaking, and the fourth one indicates if the second person is speaking.

In principle, the probability of each possible state ($\mathbf{x}_t$) in the presence of our observation ($\mathbf{y}_t$) is evaluated using Bayes' rule [29]:

$$p(\mathbf{x}_t|\mathbf{y}_t)=\frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y}_t)}=\frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{\sum_{\mathbf{x}_t}p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t)}; \quad (10)$$

unfortunately, $p(\mathbf{y}_t|\mathbf{x}_t)$ is not straightforward to compute for all different states.

In the video modality, for instance, imagine that only one face window is detected, from which we extract the corresponding observation $\mathbf{v}_t$. In that case, it is easy to compare $p(\mathbf{v}_t|\mathbf{x}_t)$ for system states with one visible person (e.g., for $\mathbf{x}_t=(1001)^T$ or $\mathbf{x}_t=(0101)^T$): It corresponds to comparing $p(\mathbf{v}_t;\boldsymbol{\theta}^V_1)$ with $p(\mathbf{v}_t;\boldsymbol{\theta}^V_2)$. In contrast, it is not clear what to do for states with two or no visible people, such as $\mathbf{x}_t=(1101)^T$ or $\mathbf{x}_t=(0001)^T$: There is no strict observation-to-parameters correspondence.

In order to solve this, the number of detected faces is added as an observation and, based on this number, the model accounts for false detection (that is, one detected region, but no visible speakers) or nondetected faces (one detected region but two visible speakers). The details of this solution are presented in Section 6.

In the audio modality, the same issue appears. It is easy to compare $p(\mathbf{a}_t(i);\boldsymbol{\theta}^A_1)$ with $p(\mathbf{a}_t(i);\boldsymbol{\theta}^A_2)$ to decide which of the two people is most likely the speaker (i.e., states like $\mathbf{x}_t=(1110)^T$ or $\mathbf{x}_t=(1101)^T$). However, it is not straightforward how to compare single-speaker states to states indicating no one speaking, or to two people speaking simultaneously, because we use person-specific (in contrast to state-specific) voice models.

A possible solution would be to model each possible combination (for instance, both of the people speaking) with a different person model. This is clearly not realistic since the number of states is exponential to the number of participants and therefore, even for a small number of people in our stream, a lot of data would be needed to obtain a good estimate for the model parameters.

Alternatively, we make a (naive, but reasonable) assumption that each person's state is independent of the other persons' states. We obtain the following factorization:

$$\begin{aligned}p(\mathbf{x}_t|\mathbf{A}_t(m))&=\prod_j p(x_t(j)|\mathbf{A}_t(m))\\&=\prod_j\frac{p(\mathbf{A}_t(m)|x_t(j))p(x_t(j))}{p(\mathbf{A}_t(m))},\end{aligned} \quad (11)$$

for which, using Bayes' Rule, we get

$$p(\mathbf{A}_t(m)|\mathbf{x}_t)=\prod_j p(\mathbf{A}_t(m)|x_t(j)), \quad (12)$$

where one is left with the difficult task of estimating $p(\mathbf{a}_t(m)|x_t(j))$ for $x_t(j)=0$, that is, model the way people affect the audio modality when silent.

The proposed framework avoids this problem with one extra variable, N^s_t, representing the number of speakers at each time slice. A GMM is trained on independent labeled data from news broadcast videos containing 1 to N speakers,[5] acquiring a generative distribution $p(\mathbf{A}_t(m)|N^s_t)$, for $N^s_t\in(0,1\ldots N)$. Note that the original training data contains a single speaker, but audio segments containing different speakers can be combined to create training data for an arbitrary number of speakers. We use the resulting GMM to evaluate the probability that an audio descriptor was generated when N^S people were speaking, and the graphical representation of this step can be seen in Fig. 3. $p(N^s_t|\mathbf{A}_t(m))$ is acquired as

$$\begin{aligned}p\big(N^s_t|\mathbf{A}_t(m)\big)&=\frac{p\big(\mathbf{A}_t(m)|N^s_t\big)p\big(N^s_t\big)}{\sum_{N^s_t}p\big(\mathbf{A}_t(m)|N^s_t\big)p\big(N^s_t\big)}\\&=\frac{p\big(\mathbf{A}_t(m)|N^s_t\big)}{\sum_{N^s_t}p\big(\mathbf{A}_t(m)|N^s_t\big)},\end{aligned} \quad (13)$$

where a uniform prior over N^s_t is assumed and this quantity is used during inference in order to avoid estimating

4. An alternative solution is to estimate the average over the frames of the stream. Experimental evaluation indicates that using the whole frame is computationally much more expensive and does not improve the results.

5. In contrast to the person voice models, which are learned from the test data.

$p(\mathbf{A}_t(i)|x_t(j))$ for $x_t(j) = 0$. The details of this procedure can be found below, in Section 6.

6 Inference

The goal of inference is to acquire the system state sequence $(\mathbf{x}^*_{1:T})$ which is the most likely given the extracted observation sequence, that is, $\mathbf{x}^*_{1:T} = \arg\max_{\mathbf{x}_{1:T}} p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$. Under the Markov assumption, the target distribution can be factorized as

$$p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = \prod_{t=0}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{y}_t|\mathbf{x}_t), \tag{14}$$

where $p(\mathbf{x}_1 = \mathbf{x}|\mathbf{x}_0)$ is the prior probability of the system being in state $\mathbf{x}$ at the first time slice. The transition probabilities $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ are taken from the factorized transition matrix $\mathcal{A}$, while $p(\mathbf{y}_t|\mathbf{x}_t)$ is the observation model.

The observation model is factorized using the person models which represent generative distributions. The observation $\mathbf{y}_t$ consists of the face descriptors acquired $(\mathbf{V}_t)$, the number of detected faces on that frame (N_t^f), the audio descriptors for that slice $(\mathbf{A}_t)$, and a discrete measure of correlation between each face and the audio stream $(\mathbf{J}_t)$. These features are independent given the system state, that is,

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t) &= p\big(\mathbf{v}_t, n_t^f, \mathbf{j}_t, \mathbf{a}_t|\mathbf{x}_t\big) = \\ &= \underbrace{p(\mathbf{v}_t|\mathbf{x}_t)\big(n_t^f|\mathbf{x}_t\big)}_{\text{Video Modality}} \underbrace{p(\mathbf{j}_t|\mathbf{x}_t)}_{\text{Joint Space}} \underbrace{p(\mathbf{a}_t|\mathbf{x}_t)}_{\text{Audio Modality}}, \end{aligned} \tag{15}$$

where the three different modalities where information comes from are indicated.

6.1 Video Modality

In the *Video Modality*, the observation is factorized into the number of detected faces, n_t^f, and the features extracted from the detected faces, $\mathbf{v}_t$. The probability of perfect face detection is set empirically to 0.9.[6] Thus,

$$p\big(n_t^f|\mathbf{x}_t\big) = \begin{cases} 0.9, & \text{if } \sum_{i=1}^{N} x_t(i) = n_t^f, \\ \dfrac{0.1}{N-1}, & \text{otherwise}, \end{cases} \tag{16}$$

where n_t^f is the number of detections returned from the face detector for frame t.

When we detect multiple faces, we need to find which face belongs to which person; all permutations are possible. A dummy variable, $\mathbf{W}_t$, is used locally and corresponds to all the possible permutations in the correspondence between person models and detected faces. That is,

$$p(\mathbf{v}_t|\mathbf{x}_t) = \sum_{\mathbf{w}_t \in \mathbf{W}} p(\mathbf{v}_t, \mathbf{w}_t|\mathbf{x}_t) = \sum_{\mathbf{w}_t} p(\mathbf{w}_t) \prod_i p\big(v_t(i); \boldsymbol{\theta}^V_{w_t(i)}\big), \tag{17}$$

where a uniform prior distribution for $\mathbf{W}_t$ is used in practice.

6. The accuracy of face detection is more dependent on the quality of the video and the viewing angle of the people's faces rather than the detection method. The 0.9 is an empirical observation of our previous research [30], [31]. As we will see in the results of Sections 8.4 and 8.5, even when this empirical estimate is inaccurate, it does not deteriorate the speaker diarization results.

Finally, the face detection procedure might not be flawless. In case $\sum_{i=1}^{N} x_t(i) > n^f$, the state $\mathbf{x}_t$ represents a state with more people than the number of detected faces. In that case, we set the probability that a person appears in the frame but is not detected by the face detector to the probability of that person being visible over the whole stream. In case of $\sum_{i=1}^{N} x_t(i) < n^f$, the remaining windows are evaluated as background, as described in Section 4.4.

6.2 Joint Space

In the joint audiovisual space, the MI feature vector is extracted from each detected face. Similarly to the video modality, nondetected people do not produce a joint space observation and we need to find which person produced each observation.

Thus, the joint space observations are evaluated in parallel to the video modality observations, and the same $\mathbf{w}_t$ settings are used:

$$p(\mathbf{j}_t|\mathbf{x}_t) p(\mathbf{v}_t|\mathbf{x}_t) = \sum_{\mathbf{w}_t \in \mathbf{W}} p(\mathbf{v}_t, \mathbf{w}_t|\mathbf{x}_t) p(\mathbf{j}_t, \mathbf{w}_t|\mathbf{x}_t) \tag{18}$$

$$= \sum_{\mathbf{w}_t} p(\mathbf{w}_t) \prod_i p(v_t(i); \boldsymbol{\theta}^V_{w_t(i)}) \prod_i p(j_t(i); \boldsymbol{\theta}^J_{w_t(i)}), \tag{19}$$

where $\boldsymbol{\theta}^J_{\mathbf{w}_t(i)}$ corresponds to the model of person $w_t(i)$.

Recall that each person model is defined in the joint audiovisual space through two sets of parameters. If $x_t^A(w_t(i)) = 1$, i.e., the system state implies that the person is speaking, then $\boldsymbol{\theta}^J_{\mathbf{w}_t(i)1}$ is used; otherwise, $\boldsymbol{\theta}^J_{\mathbf{w}_t(i)0}$—see (9) in Section 4.3.

6.3 Audio Modality

In the *audio modality*, the observation model is $p(\mathbf{a}_t|\mathbf{x}_t)$, which, since there is more than one audio descriptor per time slice, becomes $\prod_m p(\mathbf{a}_t(m)|\mathbf{x}_t)$. It is challenging to compare the probabilities of the audio descriptors when the number of speakers is different and, in particular, when this number is zero.

Introducing the random variable N_t^s, $p(\mathbf{x}_t|\mathbf{a}_t(i))$, we can avoid computing $p(\mathbf{a}_t(i)|\mathbf{x}_t(j))$ for $\mathbf{x}_t(j) = 0$ in (11) using the following factorization [32]:

$$p(\mathbf{x}_t|\mathbf{a}_t(m)) = p(N^s(\mathbf{x}_t)|\mathbf{a}_t(m)) \frac{\prod_{j:\mathbf{x}(j)=1} p\big(\mathbf{a}_t(m); \boldsymbol{\theta}^A_j\big)}{\sum_{\mathbf{x}:n_t^s(x)=n_t^s} \prod_{j:\mathbf{x}_t(j)=1} p\big(\mathbf{a}_t(m)|\boldsymbol{\theta}^A_j\big)}. \tag{20}$$

Note that $p(n|\mathbf{x}_t)$ is 1 for n, equal to the number of active speakers implied by $\mathbf{x}_t$ (denoted by $N^s(x_t)$) and 0 otherwise. Intuitively, the first decision involves what partition of the whole probability mass can be assigned to the groups of system states with the same number of speakers. Then, this mass is divided over the members of the group. The latter can be performed without explicitly modeling the generative distributions of nonspeaking participants.

6.4 Viterbi Decoding

The factorizations presented in (17), (19), and (20) combined with the factorized transition matrix $\mathcal{A}$ and the probability vector π are adequate to evaluate $p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T})$. From this, we could acquire $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ using Bayes Rule, but this is

intractable because the number of $\mathbf{x}_{1:T}$ grows exponentially to the length of the stream. However, thanks to the Markovian assumption we made, the single state sequence $\mathbf{x}^*_{1:T}$ that maximizes the likelihood of the observation sequence $\mathbf{y}_{1:T}$ can be acquired in linear time with the Viterbi algorithm [25].

7 Learning

In our approach, we assume no labeled data or prior knowledge of person models. Instead, the person models are acquired using the Expectation Maximization (EM) algorithm on the feature vectors extracted from the multimodal stream. In a nutshell, in the E-step, we compute the expectation over the variables $\mathbf{X}_t$ which represents the probability that each person is responsible for each of the observations in the audio, video, and joint audiovisual space. In the M-step, we assume that these expectations correspond to the actual state of the hidden variables and thus do not require any labeled data. Using the observations which are automatically extracted from the stream, we can then set the parameters of the each person model to the values that maximize the complete data log likelihood.[7] This allows us to segment the audiovisual sequence automatically in person-specific segments and learn the parameters of the corresponding person models. Crucially, the joint audiovisual space prevents us from having diverging audio and video models and from assigning the wrong audio model to a person's video model.

8 Experiments

The experiments were set up to test two hypotheses:

- The proposed framework incorporates video information efficiently and improves over the state-of-the-art audio-based speaker diarization.
- The framework does not require any prior knowledge and successfully incorporates the video stream in widely different scenarios.

In order to test these hypotheses, three experiments were performed, each experiment run on three different recordings. The **first experiment** evaluates the performance in speaker diarization on all the recordings using only the audio part of the proposed framework and compares it to the state-of-the-art audio-based speaker diarization system of Wooters and Huijbregts [11]. The results of this experiment will define the difficulty of each recording and serve as the baseline to measure the relative improvement by adding multimodal information.

The **second experiment** analyzes the video and joint audiovisual modality of the streams. The potential quality of the different video streams is explored to evaluate how much information can be extracted in each case—the better the analysis of the video stream, the higher its potential for speaker diarization. Moreover, we assess the quality of the observations in the joint audiovisual space which have been proposed directly for speaker diarization, e.g., [15].

The **third experiment** evaluates speaker diarization using the multimodal approach. This experiment investigates the improvement achieved using multiple modalities compared to 1) using only the audio part of the proposed model or 2) using the state-of-the-art audio-based analysis.

7. For the equations used in the M-step, see [33].

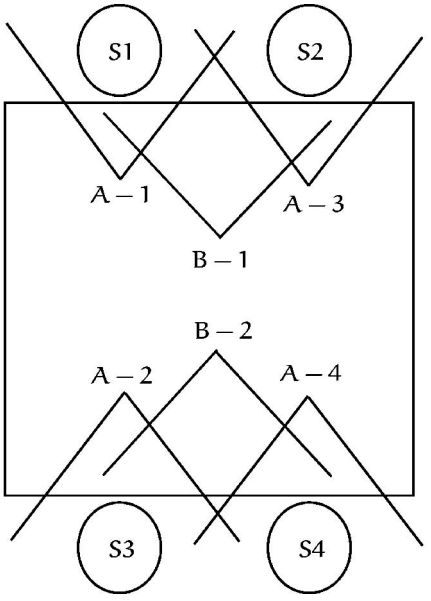


Fig. 5. The IDIAP meeting room diagram illustrates the position of the speakers $S1$-$S4$, the position of the cameras for the first setting $A1$-$A4$ and the second camera setting $B1$-$B2$. The diagram does not give an exact mapping of the equipment but approximately indicates the relative position of the different elements.

8.1 Data Sets and Performance Measure

Three different recordings are used to assess the applicability of the model in different scenarios and draw a more definite conclusion for the speaker diarization improvement when adding multimodal information. Two meeting recordings come from smart meeting rooms and they were part of the Augmented Multimodal Interaction (AMI) data set [34]. They were used in the NIST RT evaluation 2007. The third recording comes from a news broadcast and it was used in the TRECVID contest data (http://www-nlpir.nist.gov/projects/trecvid/). More specifically:

1. The first meeting recording comes from the IDIAP smart meeting room [35], which lasts approximately 30 minutes. There are four participants, seen from seven different cameras. In our experiments, two different sets of cameras were considered, and a diagram is available in Fig. 5. In set A, there is one camera for each of the participants. Frames recorded in this setting are shown in the top row of Fig. 6. In set B, seen in the bottom row of Fig. 6, there are two cameras, with two participants visible on each. In this case, each face is visible at a lower resolution.
2. The second meeting recording comes from the University of Edinburgh smart meeting room. The meeting lasts approximately 20 minutes and the four participants are visible from four cameras, which resembles the A setting of Fig. 6.
3. The third recording comes from a news broadcast video. Five people appear in the stream, but only three of them ever speak. Seven cameras were used for this recording, and each frame either corresponds to a single camera or to a combination of two cameras' parts—see Fig. 7.

Fig. 6. The upper line shows frames from setting A of the recording, while the bottom line frames from setting B.

All the recordings have high frame rate (25 fps) and well-aligned audio and video streams. The ground truth has been annotated manually with frame precision. The output of the model is a label for each frame corresponding to the identity of the speaking person(s). The speaker diarization is measured as the accuracy of this labeling:

$$\text{Overall Accuracy } = \frac{\text{Frames labeled correctly}}{\text{Total number of frames}}, \tag{21}$$

$$\text{Accuracy for Speaker } X = \frac{\text{Frames labeled correctly as speaker} X}{\text{Ground truth frames of speaker } X}. \tag{22}$$

Most audio-based approaches, including Wooters and Huijbregts [11], perform nonspeech detection and exclude the detected frames from classification. In our framework, nonspeech is one of the system states where the audio state for all the participants is 0.

8.2 Defining the Number of Speakers

The proposed framework assumes that the number of speakers is known, in contrast to the audio-based approaches submitted to RT evaluations. The methods that automatically detect the number of speakers are usually based on the Bayesian Information Criterion (BIC), defined as

$$BIC(\mathbf{Y_{1:T}}; \boldsymbol{\theta}) = \log p(\mathbf{Y_{1:T}}; \boldsymbol{\theta}) - \frac{1}{2}|\boldsymbol{\theta}| \log(n), \tag{23}$$

where $Y_{1:T}$ is all our observations, $\boldsymbol{\theta}$ all the model parameters, $|\boldsymbol{\theta}|$ is the number of free parameters of the model, and n the total number of observations. Sometimes, the weight of the penalization term $\frac{1}{2}|\boldsymbol{\theta}| \log(n)$ is adjusted by a parameter λ to control the significance of the parameter-related error term [7]. Selecting λ to perform well on a particular data set, however, is a form of overfitting.

The state-of-the-art audio-based speaker diarization method of Wooters and Huijbregts uses the ΔBIC criterion to define the final number of speakers. This criterion evaluates whether two speakers are actually the same person, i.e., if two clusters should be merged, by comparing the BIC score of both cases. In order to avoid setting the arbitrary λ parameter, the merged cluster is allowed to use twice as many free parameters as the original clusters, and, in this way, the second terms on the right-hand side of 23 cancel out.

Fig. 7. The news broadcast diagram illustrates the seven different cameras used. The rectangles next to each camera represent a timeline where black represents no signal, white means that the recording corresponds to the camera's view, and gray menas that the recording contains a part of the camera's view. In the example frames, we see, from left to right, CAM4, CAM4+CAM6, CAM6, and CAM1+CAM4.

TABLE 1
The Number of Speakers Detected from Different Criteria

	IDIAP A	IDIAP B	EDI	News
Number of Speakers	4	4	4	3
Number of Visible Persons	4	4	4	5
ΔBIC + Wooters	2	2	2	1
BIC + Audio	3	3	2	1
BIC + Multimodal	4 (4)	4 (4)	5 (5)	3 (5)

In the multimodal approach, the number in the bracket shows the total number of detected people, even if they do not speak.

TABLE 2
The Audio-Based Speaker Diarization Results

Method	IDIAP meeting	EDI Meeting	News Broadcast
Proposed audio	63%	**80**%	72%
Wooters *et al.*	**70**%	76%	**77**%
Wooters NIST	75%	82%	84%

Bold denotes the best performing method for frame-precision speaker diarization. The Wooters NIST method is applied on 2.4-second windows, which makes the problem easier, and it is provided as reference.

In the proposed framework, we cannot apply this simplification since a different number of person models will lead to a different number of parameters. Keeping the number of parameters fixed is against the principal objective of the BIC, which is to select the model with the correct complexity.

In our implementation, we choose the number of speakers that results in the highest BIC score. This is done after running each model on the whole data sequence, in contrast to ΔBIC, which is computed for consecutive segments. Table 1 lists the final number of people detected using 1) the ΔBIC criterion and the approach of Wooters and Huijbregts, 2) the BIC criterion on the audio part of our model, and 3) the BIC on the full proposed model.

8.2.1 Conclusions

Wooters and Huijbregts report that their method works well when the minimum duration of a single-speaker window is set to a large value [11]. This, indirectly, gives a smaller penalty to BIC since there is a smaller number of total points (n) to be evaluated.[8] Indeed, in our frame-precision discretization, the ΔBIC criterion leads to two speakers recognized from each of the meetings and a single speaker from the news video.

In the audio part of the proposed model, we face similar difficulties. BIC penalizes the high number of parameters required for each person model and leads to a final choice that underestimates the total number of speakers. When we add the information of the visual and joint audiovisual space, the results improve, detecting the correct number of visible persons and speakers in three out of the four recordings.

In the multimodal case, the BIC score is dominated by the video modality because the video modality produces very confident classification for each person appearing—see Section 8.4. In the Edinburgh recording, a part of the background is repeatedly detected as a face, and therefore the BIC score indicates that an extra person is beneficial since it explains all the misdetected windows.

The method of detecting the number of people in our framework favors the video modality of the data. Thus, we can draw few conclusions for its generalization properties. In a more sound approach, we can extend the proposed framework in a straightforward manner to detect the number of speakers automatically by including a prior over this number using, e.g., a Dirichlet Process. For the remainder of the experiments, all methods are provided with the correct number of people—which in the news recording is three for the audio-based models and five for the multimodal one.

8.3 Audio-Based Speaker Diarization

Here, we describe the experiment where we compare the audio part of our model to the state-of-the-art audio speaker diarization method of Wooters and Huijbregts [11]. Their method was originally applied in 2.4-second windows. Here, we apply it to both frame-duration windows, denoted with *Wooters et al.*, and 2.4-second windows, denoted with *Wooters NIST*. The former compares the two methods on the same problem, while the latter relates the results reported here to those of the NIST RT evaluation [2].

Table 2 contains the overall results of the two approaches on all recordings. In short, when the high-precision, frame-duration windows are used, Wooters and Huijbregts' method performs slightly better in the News Broadcast and IDIAP meeting and slightly worse in the Edinburgh meeting. The results of the two approaches are expected to be similar: The same features (MFCC) and parameter assumptions (GMM) are used. The differences come from the way silence and multispeaker parts are modeled.

8.3.1 Wooters' Method Results

Wooters' method performs a complex clustering of the audio descriptors, taking measures to avoid overfitting the data at hand. The optimization details used in our implementation are those suggested in the paper, which are specifically fine-tuned for the meeting videos of the contest [11]. Note that the results reported for the implementation of Wooters' method in this work differ slightly from the accuracy reported in the contest.[9] This is because of the *setup* of our work and the *scoring system* of the NIST contest.

The *setup* of this paper classifies 40-ms windows rather than the 2.4-second windows used for the contest. This is a much harder task since there is less information in every classification window. However, in tasks such as ASR, this high-temporal precision segmentation is required to produce clean, speaker-specific training data.

The contest *scoring system* evaluated excerpts of 10-12 minutes of the meetings rather than the whole meeting, but which excerpts were specifically chosen is not publicly available. In our experiments, all of the IDIAP meeting and 20 minutes of the Edinburgh meeting were used. Furthermore, the NIST evaluation does not evaluate the labeling

8. Notice that the log-likelihood scales linearly with n, while the parameter penalization scales logarithmically with it.

9. The speaker diarization accuracy for all the meeting excerpts was reported to be 79.26 percent [11].

TABLE 3
Speaker Diarization Results Using Only the Audio Modality

IDIAP MEETING					
Audio Only	NS	S1	S2	S3	S4
NS	**0.94**	0.02	0.01	0.02	0.01
S1	0.23	**0.58**	0.06	0.07	0.06
S2	0.17	0.22	**0.53**	0.05	0.03
S3	0.17	0.17	0.02	**0.60**	0.04
S4	0.10	0.13	0.03	0.06	**0.68**
Wooters	NS	S1	S2	S3	S4
NS	**0.97**	0.03	0.00	0.00	0.00
S1	0.12	**0.68**	0.07	0.08	0.05
S2	0.08	0.26	**0.49**	0.11	0.06
S3	0.13	0.10	0.08	**0.65**	0.04
S4	0.06	0.12	0.03	0.06	**0.73**

EDINBURGH MEETING					
Audio Only	NS	S1	S2	S3	S4
NS	**0.79**	0.06	0.04	0.06	0.04
S1	0.11	**0.77**	0.03	0.02	0.04
S2	0.04	0.07	**0.79**	0.02	0.06
S3	0.01	0.02	0.02	**0.87**	0.06
S4	0.01	0.07	0.04	0.05	**0.80**
Wooters	NS	S1	S2	S3	S4
NS	**0.89**	0.05	0.01	0.02	0.03
S1	0.15	**0.74**	0.02	0.02	0.04
S2	0.04	0.07	**0.79**	0.02	0.05
S3	0.13	0.01	0.01	**0.78**	0.05
S4	0.09	0.08	0.04	0.05	**0.72**

NEWS BROADCAST

Audio Only	NS	S1	S2	S3
NS	**0.89**	0.04	0.04	0.03
S1	0.06	**0.84**	0.05	0.05
S2	0.08	0.06	**0.71**	0.15
S3	0.07	0.08	0.23	**0.62**

Wooters	NS	S1	S2	S3
NS	**0.93**	0.03	0.02	0.02
S1	0.02	**0.88**	0.07	0.03
S2	0.04	0.03	**0.80**	0.13
S3	0.02	0.03	0.28	**0.67**

*Rows contain ground truth, while columns contain classification labels. Labels are **N**on**S**peech, and **S**peaker 1-4.*

accuracy in a 0.25-second collar around the speaker change points. In this way, high-precision speaker change detection and accurate labeling of short utterances is not necessary. This substantially improves accuracy results and it is in contrast to the evaluation carried out here, where every cough is taken into account.

8.3.2 Confusion Matrices

Table 3 lists the confusion matrices for each one of the recordings. Wooters' method favors the most dominant parts of the stream, for instance, Silence and Speaker 1 in the IDIAP meeting. In this way, the overall classification accuracy increases at the expense of a lower classification accuracy for persons that vocalize less. Our method performs a slightly worse but more balanced classification. Moreover, Wooters' silence detection method continuously favors classification of frames as nonspeech. In the Edinburgh meeting, which has significantly fewer silence parts, our audio model performs best.

8.4 Experiment 2: Video and Mutual Information Analysis

8.4.1 Video Analysis

Table 4 concisely presents the accuracy on *face detection* and detected window *recognition* of the each recording. This accuracy is reported based on manually labeled ground truth.[10]

Face detection is performed with the Viola-Jones face detector [26]. The Viola-Jones face detector makes two types of mistakes: 1) faces that are not detected (misses) and 2) nonfaces that are detected as faces (false positives). In the *recognition* part, no information can be extracted from a face that was not detected. In contrast, in the case of a false positive, the system must be able to robustly handle the case and classify the region as background. The recognition accuracy corresponds to the percentage of correct detections which are assigned to the right person or the percentage of background false detection which are classified as background.

In the meeting videos, the percentage of frames in which the number of faces was correctly detected is much lower than the 90 percent assumed from the model[11] (Section 6). This is because most of the meeting is spent on presentations and the participants are often seen from the side by the camera and they are not detected by the Viola-Jones face detector. The detected faces, including false positives, are classified with near-perfect accuracy. This is very important since a misclassified face window would mislead the multimodal speaker diarization.

The two different camera settings for the IDIAP meeting video provide us with varying qualities of video modality. Setting B detects fewer faces and returns many more false positives. This is because the faces are seen from a greater distance, and therefore the face detector needs to evaluate lower resolution windows.

10. See the online supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.47, for detailed results and example frames.

11. This face detection accuracy is set empirically and reflects the expected accuracy in a novel recording from unknown context.

TABLE 4
The Video Analysis Accuracy on Detection and Classification

Task	IDIAP A	IDIAP B	Edinburgh	News
Face Detection	0.63	0.52	0.63	0.99
Face Recognition	0.99	0.94	0.99	1.00
Background Recognition	0.99	0.95	0.99	1.00

Face detection performs best in the news video. This is because people are looking straight in the camera, while their face covers the largest part of the frame. The background around them is artificially generated and does not resemble a face.

8.4.2 Mutual Information Analysis

The observations of the joint space, $\mathbf{J}_t$, have a dual role. First, they indicate the correct face-to-voice correspondence. Second, they improve the speaker diarization performance by evaluating whether a detected face corresponds to a speaker or not. It is hard to evaluate the speaker diarization performance of the proposed features alone since, in many of the frames, the corresponding person is not visible. Moreover, without the audio stream observations and, more specifically, the distribution over the number of speakers at each point of the stream, the person states x_t^A. This will results in the EM splitting the observations of each person into two clusters and randomly assigning one to speaking with $x^A(n) = 1$ and one to not speaking with $x^A(n) = 0$.

In order to evaluate the quality of the information of the joint audiovisual space, we can follow the approach of Hershey and Movellan in [14]. In the two meetings with setting A, we compute the MII of each frame and make the naive assumption that the frame with the highest MI corresponds to the speaker. The corresponding speaker diarization results are presented in Table 5.

Note, in Table 5, that the naive assumption of one speaker per window essentially ignores the nonspeaker segments.

8.4.3 Conclusions

The face detection results are dependent on the choice of preprocessing method, i.e., here the Viola-Jones face detector [26]. The specific detector is known to perform very well in frontal faces and to ignore faces appearing rotated or seen from the side. The classification accuracy of the detected faces is nearly perfect. The few mistakes occur in cases where a face is lost temporarily while, at the same time, a false positive window appears in the frame. In such cases, the transition probability favors classifying this false positive as the missing person. Further processing of the video modality can eliminate these false positives.

The mutual information contains information relative to speaker diarization. Using a naive assumption, it produces results correlated to the active speaker and, incorporated under a sound probabilistic model, it will improve the overall results of speaker diarization.

TABLE 6
The Overall Speaker Diarization Accuracy Achieved by Different Input Modalities

Method	IDIAP A	IDIAP B	Edinburgh	News
Wooters *et al.*	70%	70%	76%	77%
Audio Only	67%	67%	80%	72%
Multimodal	**84**%	**77**%	**89**%	**94**%

8.5 Experiment 3: Multimodal Speaker Diarization

The high accuracy in the visual analysis allows us to integrate the speaker diarization information of the joint audiovisual space efficiently. The overall speaker diarization accuracy results for the multimodal approach are reported in Table 6, where the multimodal approach clearly outperforms the single modality analysis.

Table 7 presents the multimodal speaker diarization results acquired in different scenarios and camera settings. The audio modality alone, presented in Section 8.3, does not distinguish between different speakers very well and the addition of the other modalities improves the results for each speaker. Note that in the multimodal approach, no speaker is favored specifically, but a similar accuracy is achieved for all of them.

Multimodal analysis produces interesting results in the multispeaker parts of the stream. In frame-precision annotation, there are many frames where multiple people vocalize. These frames are commonly labeled as a single person [2]. In our work, they are treated as multispeaker segments and they are accounted as a correct classification only in case the correct multispeaker label is selected. In short, the multimodal approach gets and average of

TABLE 5
Speaker Diarization Results Using Only the MI of Each Frame in Setting A

	IDIAP MEETING					EDINBURGH MEETING				
MI Only	S1	S2	S3	S4	**MI Only**	S1	S2	S3	S4
NS	0.29	0.31	0.15	0.25	NS	0.24	0.33	0.33	0.08
S1	**0.33**	0.20	0.17	0.30	S1	**0.29**	0.20	0.13	0.37
S2	0.15	**0.48**	0.12	0.25	S2	0.16	**0.31**	0.23	0.30
S3	0.19	0.26	**0.24**	0.31	S3	0.19	0.27	**0.27**	0.34
S4	0.19	0.27	0.21	**0.33**	S4	0.07	0.39	0.15	**0.39**

Rows contain ground truth, while columns contain classification labels.

TABLE 7
Confusion Matrix for Multimodal Speaker Diarization Results

MULTIMODAL SPEAKER DIARIZATION

IDIAP A	NS	S1	S2	S3	S4
NS	**0.84**	0.06	0.03	0.03	0.04
S1	0.02	**0.82**	0.01	0.07	0.08
S2	0.01	0.03	**0.76**	0.10	0.10
S3	0.01	0.08	0.07	**0.77**	0.08
S4	0.02	0.02	0.09	0.02	**0.85**

IDIAP B	NS	S1	S2	S3	S4
Non-Speech	**0.88**	0.05	0.01	0.05	0.01
S1	0.02	**0.82**	0.01	0.07	0.08
S2	0.11	0.06	**0.68**	0.14	0.01
S3	0.01	0.07	0.06	**0.79**	0.04
S4	0.06	0.04	0.16	0.09	**0.64**

Edinburgh	NS	S1	S2	S3	S4
NS	**0.84**	0.05	0.03	0.05	0.03
S1	0.08	**0.84**	0.03	0.02	0.03
S2	0.03	0.06	**0.84**	0.02	0.05
S3	0.01	0.02	0.02	**0.91**	0.04
S4	0.01	0.05	0.03	0.04	**0.87**

News	NS	S1	S2	S3
NS	**0.96**	0.00	0.00	0.04
S1	0.02	**0.89**	0.06	0.02
S2	0.02	0.04	**0.91**	0.03
S3	0.02	0.02	0.01	**0.95**

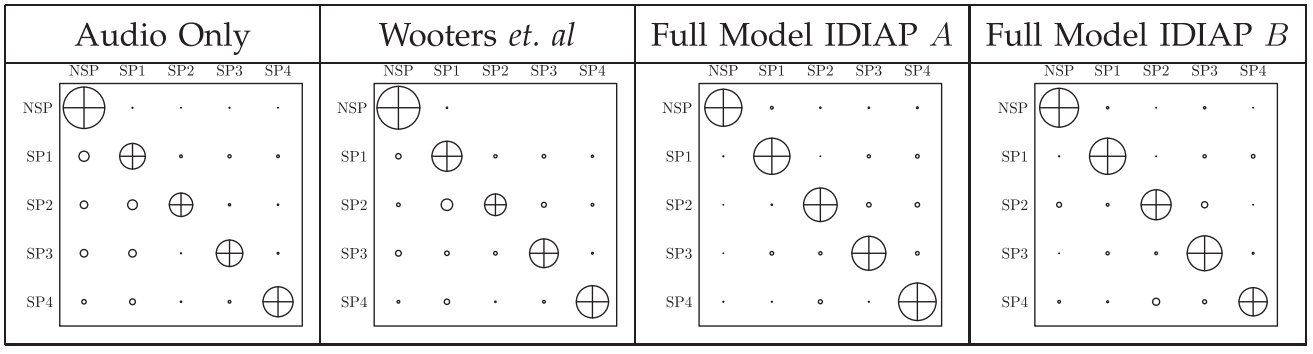


Fig. 8. A visual representation of the speaker diarization results for the IDIAP recording.

71 percent accuracy in contrast to just 28 percent for the audio-based models.[12]

8.6 Significance of the Results

The multimodal approach produces higher speaker diarization accuracy than the audio-based approaches. We perform a t-test to evaluate the statistical significance of this difference. The comparison between the proposed approach and the results of Wooters gives a t-value of 6.0812, which means that the results are significantly different at a confidence level of 99 percent. The results between the proposed model and the audio part alone produce a t-value of 4.752, which implies statistical significance with a confidence level of 98 percent. Comparison between the audio model and Wooters' method gives a t-value of 0.8866, which corresponds to no statistically significant difference.

9 Discussion

The multimodal approach beats the results of the method proposed by Wooters and Fuijbregts [11] in terms of speaker diarization in the experimental data. This was expected since more information is used. The final results of 84 and 89 percent in the IDIAP and Edinburg recordings beat the 79 percent state-of-the-art performance reported in the RT benchmark [2] under a much more difficult objective: We classify windows with 40-ms precision instead of 2.4 seconds. This high precision speaker diarization is essential for automatic transcript generation and more useful automatic speech recognition [36].

Furthermore, the method of Wooters and Huijbregts assumes a single speaker per window. On one hand, since in a recording there are windows where multiple speakers vocalized together, these windows cannot be classified correctly by Wooters' method, leading to lower classification accuracy. On the other hand, these windows are a very small part of the stream. A model which considers all possible speaker combinations, such as the one we propose, has a much harder task in the remaining, major part of the stream.

In Figs. 8 and 9, the experimental results are represented graphically. In these plots, there is one circle for each element of the confusion matrix with a radius proportional to the corresponding element. A perfect classification has large circles on the diagonal—nondiagonal circles represent misclassification. It is clear that the proposed multimodal

12. See supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.47, for detailed results on the multiperson parts.

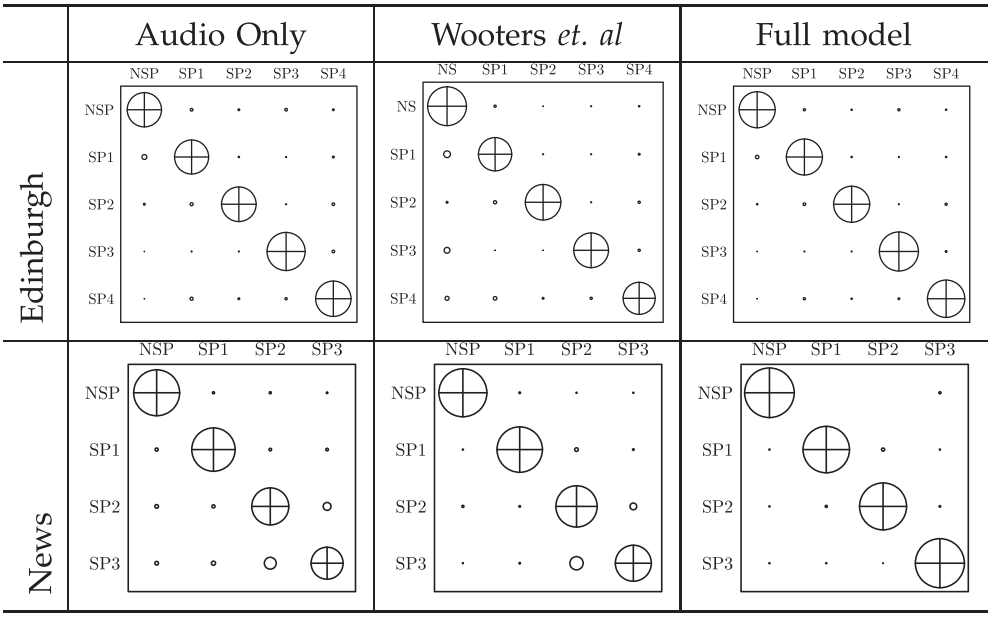


Fig. 9. Visual representation of speaker diarization for the Edinburgh and news videos.

approach performs equally well over all speakers, even those that vocalize less, because it can incorporate the video stream, where speakers and nonspeakers are equally represented. In contrast, the audio-based analysis will inevitably focus on cases dominant in the audio space, which were the speaker 1 and the nonspeech segments in the IDIAP meeting and Speaker 3 and nonspeech segments in the Edinburgh meeting.

10 Conclusions

This paper presented a probabilistic framework that performs probabilistic multimodal speaker diarization. We have shown that:

1. The model incorporates all the available sources of information and outperforms the current state of the art in single modality analysis.
2. The proposed framework relaxes the assumptions about the position of the recording equipment. Multiple microphones can be merged into one channel, while multiple camera views can be used as long as the same person does not appear twice.

Moreover, improvements over previous approaches are that:

1. The Bayesian nature of the model incorporates the temporal information of the data and performs parallel speaker segmentation and clustering directly on the test recording.
2. The parts of the stream in which two or more people speak simultaneously are treated under the same framework. This allows both to detect the correct speaker(s) and to avoid using such parts to learn a single-speaker's model.

The framework is robust and it provides high-accuracy speaker diarization results in a variety of scenarios and camera settings. The proposed fusion method proves very effective since incorporating the video modality improves the results on all the recordings. Moreover, the probabilistic nature of the proposed approach allows straightforward incorporation of further modalities, features, or prior knowledge.

In terms of the *features* proposed here, we do not claim optimality in speaker diarization accuracy. Different feature choices, voice, or appearance models could improve the results significantly. However, future developments in high-quality features for speaker diarization, such as, perhaps, lip-reading or motion detection features, can be incorporated under the proposed model. The only prerequisite is to find a suitable probability distribution over these features, conditioned on the identity of the speaker.

References

[1] D.A. Reynolds and P.A. Torres-Carrasquillo, "Approaches and Applications of Speaker Diarization," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 953-956, 2010.

[2] J.G. Fiscus, J. Ajot, and J.S. Garofolo, "The Rich Transcription 2007 Meeting Recognition Evaluation," *Multimodal Technologies for Perception of Humans*, pp. 373-389, Springer-Verlag, 2008.

[3] R. Bakis, S. Chen, P. Gopalakrishnan, R. Gopinath, L. Polymenakos, and M. Franz, "Transcription of Broadcast News Shows with the IBM Large Vocabulary Speech Recognition System," *Proc. Speech Recognition Workshop*, pp. 67-72, 1997.

[4] J. luc Gauvain, L. Lamel, and G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. Int'l Conf. Spoken Language Processing*, pp. 1335-1338, 1998.

[5] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," *Proc. Int'l Speech Comm. Assoc. Workshop Speaker Recognition: A Speaker Oddyssey*, 2001.

[6] M. Yamaguchi, M. Yamashita, and S. Matsunaga, "Spectral Cross-Correlation Features for Audio Indexing of Broadcast News and Meetings," *Proc. Ninth European Conf. Speech Comm. and Technology*, pp. 613-616, 2005.

[7] P. Delacourt, D. Kryze, and C.J. Wellekens, "DISTBIC: A Speaker-Based Segmentation for Audio Data Indexing," *Speech Comm.*, vol. 32, pp. 111-126, 2000.

[8] S.S. Chenn and P. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 645-648, 1998.

[9] K. Mori and S. Nakagawa, "Speaker Change Detection and Speaker Clustering Using VQ Distortion for Broadcast News Speech Recognition," *Systems and Computers in Japan,* vol. 34, pp. 413-416, 2001.
[10] R. Gangadharaiah, B. Narayanaswamy, and Narayanaswamy, "A Novel Method for Two-Speaker Segmentation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* 2004.
[11] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," *Multimodal Technologies for Perception of Humans,* pp. 509-519, Springer-Verlag, 2008.
[12] Z. Barzelay and Y.Y. Schechner, "Harmony in Motion," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2007.
[13] E. Kidron, Y.Y. Schechner, and M. Elad, "Pixels That Sound," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 88-95, 2005.
[14] J. Hershey and J. Movellan, "Using Audio-Visual Synchrony to Locate Sounds," *Advances in Neural Information Processing Systems,* vol. 12, pp. 813-819, MIT Press, 1999.
[15] T. Darrell, J.W. Fisher III, and P. Viola, "Audiovisual Segmentation and the Cocktail Party Effect," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 32-40, 2000.
[16] H.J. Nock, G. Iyengar, and C. Neti, "Multimodal Processing by Finding Common Cause," *Comm. ACM,* vol. 47, no. 1, pp. 51-56, 2004.
[17] G. Iyengar, H.J. Nock, and C. Neti, "Audio-Visual Synchrony for Detection of Monologues in Video Archives," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 329-332, 2003.
[18] J.W. Fisher III and T. Darrell, "Probabilistic Models and Informative Subspaces for Audiovisual Correspondence," *Proc. European Conf. Computer Vision,* Part III, pp. 592-603, 2002.
[19] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. wei He, A. Colburn, Z.Z.Z. Liu, and S. Silverberg, "Distributed Meetings: A Meeting Capture and Broadcasting System," *Proc. 10th ACM Int'l Conf. Multimedia,* 2002.
[20] Y. Chen and Y. Rui, "Real-Time Speaker Tracking Using Particle Filter Sensor Fusion," *Proc. IEEE,* vol. 92, no. 3, pp. 485-494, Mar. 2004.
[21] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple Person and Speaker Activity Tracking with a Particle Filter," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* May 2004.
[22] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal Multispeaker Probabilistic Tracking in Meetings," IDIAP-RR 66, IDIAP, 2004.
[23] M.J. Beal, H. Attias, and N. Jojic, "Audio-Video Sensor Fusion with Probabilistic Graphical Models," *Proc. European Conf. Computer Vision,* 2002.
[24] Z. Ghahramani, M.I. Jordan, and P. Smyth, "Factorial Hidden Markov Models," *Machine Learning,* MIT Press, 1997.
[25] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, Feb. 1989.
[26] P. Viola and M. Jones, "Robust Real-Time Object Detection," *Proc. Second Int'l Workshop Statistical and Computational Theories of Vision—Modeling, Learning, Computing, and Sampling,* 2001.
[27] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual Categorization with Bags of Keypoints," *Proc. European Conf. Computer Vision Int'l Workshop Statistical Learning in Computer Vision,* 2004.
[28] T. Darrell, J.W. Fisher III, W.T. Freeman, and P. Viola, "Learning Joint Statistical Models for Audio-Visual Fusion and Segregation," *Advances in Neural Information Processing Systems,* vol. 13, pp. 772-778, MIT Press, 2000.
[29] B. Thomas, "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Trans. Royal Soc.,* vol. 53, pp. 370-418, 1763.
[30] A.K. Noulas, N. Vlassis, and B.J.A. Kröse, "Cross Entropy for Learning in Multi-Modal Streams," *Proc. Joint Workshop Multi-Modal Interaction and Related Machine Learning Algorithms,* 2007.
[31] A.K. Noulas and B.J.A. Kröse, "EM Detection of Common Origin of Multi-Modal Cues," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 201-208, 2006.
[32] A.K. Noulas and B.J.A. Krose, "A Hybrid Generative-Discriminative Approach to Speaker Diarization," *Proc. Fifth Int'l Workshop Machine Learning for Multimodal Interaction,* pp. 98-109, 2008.
[33] A. Noulas, "Audiovisual Fusion for Speaker Diarization," PhD dissertation, Univ. of Amsterdam, 2010.
[34] J. Carletta, "Announcing the AMI Meeting Corpus," *The ELRA Newsletter,* vol. 1, no. 1, pp. 3-5, Jan.-Mar. 2006.
[35] D.C. Moore, "The IDIAP Smart Meeting Room," IDIAP, IDIAP-COM 07, 2002.
[36] X. Anguera, C. Wooters, and J. Hernando, "Automatic Cluster Complexity and Quantity Selection: Towards Robust Speaker Diarization," *Proc. Third Joint Workshop Multimodal Interaction and Related Machine Learning Algorithms,* pp. 248-256, 2006.

Athanasios Noulas received the PhD degree from the University of Amsterdam, where he focused on audiovisual fusion for speaker diarization. Currently, his research interest is information fusion in time series data and, more specifically, its application to financial markets.

Gwenn Englebienne received the PhD degree in computer science from the University of Manchester, United Kingdom, on jointly modeling the audio and video of talking heads, for which he received the department's Best Thesis award. He is a postdoctoral researcher at the University of Amsterdam, where he focuses on machine learning with Bayesian probabilistic methods. His research interests are in the probabilistic modeling of human behavior and interaction using various modalities, such as audio, video, and simple binary sensors.

Ben J.A. Kröse is a professor at the University of Amsterdam and at the Amsterdam University of Applied Science. His research focuses on interactive smart devices, which is expected to be widely applied for smart services in health, safety, well-being, security, and comfort. He is scientific manager of Create-IT, a research center for IT and the creative industry. In the fields of intelligence and autonomous systems, he has published 33 papers in scientific journals, edited five books and special issues, and has published more than 100 conference papers. He owns a patent on multicamera surveillance. He is a member of the IEEE, Dutch Pattern Recognition Association, and Dutch AI Association.