11

A Review of Recent Advances in Speaker Diarization with Bayesian Methods

Themos Stafylakis[1] and Vassilis Katsouros[2]
[1]*Institute for Language and Speech Processing, "Athena"R.C. & National Technical University of Athens, Department of Electrical and Electronic Engineering*
[2]*Institute for Language and Speech Processing, "Athena" R.C.*
Greece

1. Introduction

This chapter aims to present some of the recent Bayesian approaches to speaker diarization (SD). SD is the task of grouping an audio document into homogenous regions, where each region should ideally correspond to the complete set of utterances that belong to a single speaker. Rich transcription, speaker adaptation of speech recognition systems and speaker recognition are some of the applications that require such a clustering procedure. Broadcast News, meeting, and telephone conversations are the main domains that SD is applied to.

SD is a fully unsupervised clustering task. Not only we are not allowed to use any target-speaker enrollment data to detect the target speakers through the acoustic stream, but the number of speakers should be considered as an unknown, too. Moreover, text-independence should also be assumed, meaning that no transcript is available, either.

Despite the effectiveness of several approaches and frameworks that have been proposed and tested in literature, the most natural and systematic approach to SD is to treat it as a model's order selection task. Once the order is estimated (i.e. the number of speakers) the task reduces to a familiar (but not trivial at all) machine learning task where the latent variables (i.e. the speaker indicators of each utterance) of given cardinality should be estimated from the observations. Therefore, a major issue we deal with is how to assess the number of speakers in a way that is simultaneously robust and efficient.

Bayesian machine learning is a highly principled paradigm and can naturally tackle model selection problems. It does so by applying consistently the rules of probability in order to infer the desired quantities, including the order of the model. Its superiority over the frequentistic statistical framework (e.g. Maximum Likelihood estimates, Classical Hypothesis testing) or semi-Bayesian approaches (e.g. MAP estimation, penalized maximum likelihood criteria) in model selection, averaging and density estimation has been verified in most (if not all) of the speaker related tasks, including identification and verification.

Several drawbacks however still exist, most of which stem from the intractability of the majority of the ideal Bayesian solutions. Many well known and effective machine learning tools cannot be applied or require severe adaptation that may drastically increase their computational complexity. Nevertheless, the introduction of powerful approximate inference method (e.g. Variational Bayes, Expectation Propagation), novel Markov-Chain Monte Carlo techniques, along with the rapid development of the Bayesian nonparametric models

(Infinite-HMMs, Dirichlet process mixture models, a.o.) allows us to create new approaches that are based on the statistical coherency of the Bayesian framework.
The rest of the chapter is organized as follows. In Section 2, some of the non- and semi-Bayesian approaches to SD is reviewed, along with some definitions and general algorithms strategies. In Section 3, the basic theory of Variational Bayes inference is presented with emphasis on mixture models, while a Variational Bayes algorithm that uses supervectors is examined in Section 4. In Section 5, we consider the use of infinite models to SD, while some ideas about further applications of Bayesian inference in diarization are discussed in Section 6. Finally, an introduction to some novel features that are utilized in speaker verification and recently in diarization are presented in the Appendix.

2. Short overview of speaker diarization approaches

In this section, a brief introduction to SD is presented, followed by some approaches that have been proposed in literature. We will refer to several algorithmic approaches and discuss some of their strengths and weaknesses. For a more complete overview of these methods we refer to (Tranter & Reynolds, 2006).

2.1 Front-end features and preprocessing steps

Before we examine the several algorithmic approaches, let us review some aspects that are common to all systems. The majority of SD systems use Mel-Frequency Cepstral Coefficients (MFCC) as front-end features, although other feature spaces have been proposed, such as Linear Frequency Cepstral Coefficients (LFCC) and Perceptual Linear Predictive (PLP), (Hermansky et al., 1985). Some systems utilize prosodic features to augment the cepstral representation (see Friedland et al. (2009)) while other approaches attempt to fuse several spaces and increase the diarization accuracy, (Gupta et al., 2007). Depending on the application field, one may consider techniques to normalize the MFCC stream, (Pelecanos & Sridharan, 2001), (Xiang et al., 2002), (Hermansky et al., 1992). These techniques aim to remove the linear channel effect and possibly the additive noise introduced by the recording chain, and are compulsory when a speaker may speak with more than one recording chains. In SD, such techniques may not be necessary; a standard assumption is that each speaker speaks only under identical conditions, i.e. recording equipment and background noise. Moreover, since the channel is unknown, these techniques unavoidably remove information that is related to the speaker and therefore increase the similarity between different speakers.
In the multiple-microphone setting (e.g. meetings), two are the main approaches. The first is to apply acoustic array processing techniques (i.e. beamforming) in order to mix the signals into a unique enhanced signal, (Anguera et al., 2007). A second approach is to utilize the estimated direction-of-arrivals (DOA) and fuse spatial and cepstral information, (Pardo et al., 2007). In our review, we will focus on the former approach when multiple microphones are in-hand.
A second step that is common to most of the algorithms is Speech Activity Detection (SAD). Silent regions of duration more that 200ms should be detected and removed from the steam. The official scoring method of NIST, the Diarization Error Rate (DER), penalizes false alarm and missed detection rates linearly. A common approach to detect speech is to assume that speech and silence follow a normal distribution each, in the log-energy domain. An Expectation Maximization (EM) algorithm with two Gaussian components is then applied, using the log energy as features. The energy feature stream is calculated using sliding windows of typically 30ms duration, with 20ms overlap, so that it is aligned with the MFCC

stream. Temporal smoothing techniques are then applied on the binary labels to discard regions of less that 200*ms* duration. Hidden Markov Model (HMM)-based EM may also be considered as well, in order to avoid the need of ad-hoc or morphological filtering techniques. Apart from the energy, periodicity based methods have been proposed. These methods utilize the facts that vowels exhibit strong (quasi-)periodicity and apply it to discriminate speech from silence. Periodicity based approaches are usually more robust to noisy environments, however they require more computational effort than the energy-based ones, (Ishizuka et al., 2010).
Finally, in the Broadcast News field, most systems discriminate between acoustic classes like speech, music, music and speech, and silence. To do so, supervised learning techniques are applied. Each class is modeled with a GMM with 128 or 256 diagonal Gaussian components using labeled training data. During the classification stage, regions that are classified as non-speech are removed from the stream, after a proper temporal smoothing on the class-label domain.

2.2 General algorithmic approaches

After the preprocessing steps described above, SD algorithms diverge into two main directions. Those that apply segmentation to the MFCC stream, which might be uniform or based on the speaker change detection algorithms (see (Chen & Gopalakrishnam, 1998)), and those that do not apply such segmentation. Following the terminology of (Meignier et al., 2006) we will refer to the former branch as step-by-step algorithms, while to the latter as integrated algorithms. Both algorithmic approaches exploit a certain characteristic that the speaker labels exhibit, which is the temporal continuity. To realize the minimum range of this continuity, note that a speaker's turn lasts no less than 1 or 2*s* while the MFCC rate is 10*ms*, typically. Step-by-step algorithms exploit this continuity in order to turn the problem into a typical unsupervised clustering task. They represent each segments using a statistical model (a single Gaussian or a GMM) and they apply clustering techniques to group them into speakers. On the contrary, the integrated algorithms exploit the temporal continuity by assuming that the transitions between speakers follow a stochastic process which can be modeled by a (first-order and time-independent) Markov chain, where the probability of self-transition is significantly greater than the one of departing from the current state. Since the labels are not directly observed (in fact, they are the desired quantities) an observation model should be added, to link each distinct label (or state) with the observations. The overall model is therefore a HMM, where the observation model (i.e. the state-emission probabilities) is usually a GMM for each state, that is capable of capturing the multimodality of the state-conditional distribution.

2.3 Distance-based and model-based approaches to speaker clustering

However, what restrain us from using standard clustering or HMM techniques is the lack of knowledge regarding the number of speakers, say K. If we a priori knew K then we would apply an EM-algorithm to learn both the model and the latent variables (i.e. the label of each MFCC frame). [1] Two are the main approaches to deal with this issue. The first approach, which is extremely common to step-by-step algorithms, is to apply agglomerative hierarchical clustering (AHC) to merge those segments being close enough, in a statistical sense. What is required is a measure of similarity (or equivalently dissimilarity) and usually a predefined

[1] This is partially true however; phoneme rate, pitch, intensity and other emotional variations that speakers may exhibit during their speech may cause failure even in this setting.

threshold. We refer to these approaches as distance-based. Most step-by-step approaches use a two stage AHC procedure; at the first stage the segments (and consequently the clusters) are modeled using a single Gaussian of full covariance matrix, while GMMs are deployed only on the next stage, to merge those clusters that had not been merged during the first step. Several of the similarity measures that are used in the second stage are discussed in the Appendix, along with the MAP-EM algorithm that is applied to train the GMMs. Note also that several hybrid algorithms exist as well. For example, the highly robust and tuning-free approach proposed in (Ajmera & Wooters, 2003) uses a uniform segmentation stage and applies a Viterbi re-segmentation algorithm each time a pair of clusters is merged. Finally, several other alternative to AHC algorithms have been proposed, including Self-Organized Maps, Spectral Clustering and the Mean-Shift algorithm, that produce competing or better SD results, (Lapidot, 2003), (Ning et al., 2006), (Stafylakis et al., 2010b).

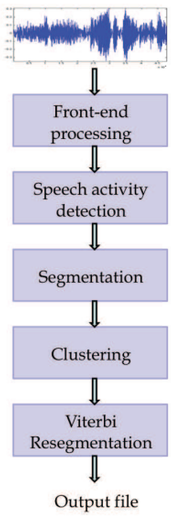

Fig. 1. Flow-chart of a baseline step-by-step algorithm

The main problem, however, regarding the distance-based approaches is their heuristic nature, in the sense that they do not propose a method to score overall clustering hypotheses. Note that the distance-based category of approaches may include even methods that rely on similarity measures that are derived from model-selection. For example, the local-Bayesian Information Criterion (BIC) ((Zhu et al., 2005), (Barras et al., 2004)) might be a model-based dissimilarity measure, however it does not correspond to the difference between scores of competing clustering hypotheses, (Stafylakis et al., 2010a). A desired property of a clustering algorithm is to be capable of providing a score to every single possible configuration of the latent variables. This is the essence behind the model-based approaches, (Fraley & Raftery, 1998). A model-based approach may be applied to a broad range of algorithms, which includes the AHC as well. To do so, we need to consider the dissimilarity between any pair of segments (or clusters of segments) as the increase or decrease of the overall score caused by the action of merging this pair. The global and the segmental settings of BIC are such examples, (Stafylakis et al., 2010a).

2.4 Penalized likelihood and its limitations

The most significant gain, however, from using model-based approaches is that it allows us to make use of the most natural and powerful tool of learning with missing data, that is the broad

class of EM algorithms, (Dempster et al., 1977), (Amari, 1995). The integrated approaches typically use an evolving HMM (E-HMM), that is an HMM with increasing number of states. For each number of states, the Viterbi of Baum-Welch algorithm is deployed to learn the latent variables and estimate the emissions. To estimate the true number of speakers and the corresponding clustering hypothesis, an appropriate model selection method is compulsory. In the step-by-step approach, a form of EM algorithm can be applied instead of the AHC algorithm, over the range of a priori plausible number of speaker and a model selection method is required in order to select amongst them, (Mackay, 2003).

The penalized likelihood criteria have become very popular, for two main reasons. First of all, they can be used to apply model selection without altering the non- or semi-Bayesian way we estimate the parameters of the model with missing data. For example, in the E-HMM approach to SD, one may use the standard Maximum Likelihood (or MAP estimate) and penalize it according to the well-known BIC penalty term. The second reason is that under some regularity conditions they are limits of the desired Bayesian quantity; the *marginal likelihood* of the model. The BIC is derived by approximating the *marginal likelihood* of the model with the Laplace method, and discarding those terms that do not scale with the number of observations.

However, there are certain drawbacks regarding this semi-Bayesian approaches. For example, there are several models for with the consistency of the BIC has not been proven. This includes all the mixture models, including GMMs and HMMs as well. Even though in cases where the regularity conditions hold, the Laplace approximation is usually inaccurate for small sample sizes. Moreover, a MAP estimate is still point estimate, since the uncertainty about the estimate is being ignored, (Mackay, 2003). Finally, many of the powerful Bayesian tools, like the use of explicit priors or the use of hierarchies to tie several parameters cannot be combined in a profound way with the BIC approximation. Therefore, it becomes evident that a fully-Bayesian treatment of SD is required, which is the objective of the rest of this chapter.

3. Methods based on Variational Bayes approximate inference

In this chapter, the use of a fully Bayesian framework to SD is examined. The term Variational Bayes (VB) refers to a set of methods (the most popular of which being the *mean-field* VB) that approximate the desired quantities (e.g. marginal likelihoods, posterior probabilities, predictive densities) by bounding the marginal likelihood of the model from below. The use of VB is SD has been pioneered by F. Valente (Valente, 2005) and has been refined by P. Kenny *et al.* (Kenny et al., 2010) by applying it to i-vectors. We should emphasize that VB is a *general purpose* (approximate) inference method and its use is not limited to finite mixture models. On the contrary, it can be applied to nonparametric models, too (e.g. Dirichlet Process Mixture Models, (Blei & Jordan, 2005)).

3.1 Fundamentals of Variational Bayes

Let as consider a family of nested models $\mathcal{M}$ and let K denote the order of the model (e.g. the number of components of a GMM, the number of states of an HMM, etc.). Let the parameter space be denoted by Θ while the set of latent variables by X. The most probable order of the model given Y is the one that maximizes $P(K|Y) \propto p(Y|K)P(K)$. Assuming uniform prior over the hypothesis space (i.e. $P(K) \propto 1$), we need to maximize the *marginal likelihood* of the model with respect to (w.r.t.) K, i.e.

$$p(Y|K) = \int p(Y, X, \Theta) dX d\Theta \tag{1}$$

Alike BIC and other Laplace approximation based approaches, the VB framework defines a lower bound of (1). It does so by (i) introducing the variational posterior $q(X,\Theta)$ (the conditioning on Y is kept implicit) and (ii) applying the Jensen inequality, as follows

$$\log p(Y|K) = \log \int \frac{p(Y,X,\Theta)}{q(X,\Theta)} q(X,\Theta) dX d\Theta \geq \int \log \left(\frac{p(Y,X,\Theta)}{q(X,\Theta)} \right) q(X,\Theta) dX d\Theta \quad (2)$$

The bound that (2) defines is known as the (negative) Variational free energy $\mathcal{F}_K(q(X,\Theta))$, while the difference between $\log p(Y|K)$ and $\mathcal{F}_K(\cdot)$ is equal to $D_{KL}(q(X,\Theta)||p(X,\Theta|Y))$. However, no further improvement can be attained without making some assumptions about the functional form of $q(X,\Theta)$. The mean field VB pretends that $X|Y$ and $\Theta|Y$ are independent, and therefore assumes that $q(X,\Theta)$ admits a factorization of the form $q(X,\Theta) = q(X)q(\Theta)$. We say so, since this factorization is only a priori possible. A posteriori, the observation of Y induces an (at least weak) correlation between X and Θ. However, this independence assumption allows as to make the optimization problem tractable by applying calculus of variations.

3.2 The VB-EM algorithm

By maximizing $\mathcal{F}_K(q(X)q(\Theta))$ w.r.t. to $q(X)$ and $q(\Theta)$ we end up with the VB-EM algorithm described below

$$\text{VB-E step: } q(X) = \frac{1}{Z_X} e^{<\log p(Y,X|\Theta)>_{q(\Theta)}} \quad (3)$$

$$\text{VB-M step: } q(\Theta) = \frac{1}{Z_\Theta} e^{<\log p(Y,X|\Theta)>_{q(X)}} p(\Theta|K) \quad (4)$$

where $< a >_b$ denotes the expected value of a w.r.t b, while Z_X and Z_Θ are the corresponding normalizing constants. Note that the existence of $p(\Theta|m)$ at the M-step induces no asymmetry between X and Θ; the prior of X is incorporated through the complete-data likelihood $p(Y,X|\Theta) = p(Y|X,\Theta)p(X|m)$.

The severe distinction between ML-EM (or MAP-EM) and VB-EM is that while the former proceeds with simple point masses $\delta(\Theta,\hat{\Theta})$ placed at the ML or MAP estimates of Θ, VB-EM captures the uncertainty in these estimates, through the posterior distribution of Θ. Each estimate of X is obtained by averaging w.r.t. to the posterior of Θ, and not by $\delta(\Theta,\hat{\Theta})$. Furthermore, the benefits from using such a fully probabilistic approach are not restricted to obtaining much richer inferences about Θ and X. Contrary to ML- and MAP-EM, VB-EM aims to maximize the marginal likelihood of models, which is the key quantity in assessing K. No penalty term is required; the marginal likelihood is all we need to obtain in order to select between the rival models.

However, we should re-emphasize that the quantity being maximized by VB is $\mathcal{F}_K(q(X,\Theta))$ and not $\log p(Y|K)$. We saw that the difference between the two terms is equal to $D_{KL}(q(X|K)q(\Theta|K)||p(X,\Theta|Y,K)) > 0$ which increases with K. Therefore, the approximation of $\log p(Y|K)$ by $\mathcal{F}_K(q(X,\Theta))$ induces a systematic bias towards simpler models and therefore VB may underestimate the true number of speakers.

3.3 Hyperparameters: centering and strength

So far, we have assumed that the hyperparameters (i.e. the variables that parametrize the prior) remain fixed during the VB-EM. Let us denote the set of hyperparameters by H. By restricting ourselves to the *conjugate* family of priors, the hyperparameters can be distinguished into two sets $H = [H^c, H^s]$. Those that parametrize *expected values* of elements of

Θ (the prior centers, denoted by H^c) and those that determine the *amount of virtual observations* carried into the prior, also known as the *strength* of the prior (e.g. the relevance factor r in (42)), denoted by H^s. Priors with large strength are called *informative*, in the sense that their impact on the posterior is significant, at least when dealing with small or medium T. In cases where only vague, unreliable, or no information at all is available about Θ, a good strategy is to keep the prior as *non-informative* as possible. Jeffreys' priors, defined as follows

$$p^J(\theta) \propto |\mathcal{I}_\Theta(\theta)|^{1/2} \tag{5}$$

where $\mathcal{I}_\Theta(\theta)$ denotes the Fisher information matrix and θ an element of Θ, are flat in the sense that they place equal probability mass on each *natural volume element* of the statistical manifold, (Snoussi, 2005). They are also limits of conjugate priors, defined as below

$$p(\theta|h_\theta^c, h_\theta^s) \propto |\mathcal{I}_\Theta(\theta)|^{1/2} \exp\left(-h_\theta^s D_{KL}(h_\theta^c||\theta)\right) \tag{6}$$

by letting the strength go to zero. However, they are rarely *proper*, since $\int |\mathcal{I}_\Theta(\theta)|^{1/2} d\theta$ usually goes to infinity, and therefore inadequate for the model selection task. Hence, one may use conjugate priors and place h_θ^s equal to the minimum value (or the minimum integer) for which (6) is proper.

A further issue regarding the strength of the prior is whether the overall amount of virtual observations should remain fixed or be allowed to vary with K. For example, the standard penalty term of BIC implies a strength that remains fixed. Hence, the more parameters we add to the model, the less informative the (implied) prior will be for each single parameter. However, this strategy can be restrictive for models having parameters whose prior requires a minimum amount of strength to be proper (e.g. covariance matrices). In such cases, this strategy bounds from above the overall number of parameters that can be used and, consequently, the number of clusters. On the contrary, letting the strength grow with the number of parameters can cause overestimation of the true order of the model.

In any case, if we choose to optimize the hyperparameters, a straightforward solution is to solve the following maximization problem

$$H^{(i+1)} = \arg\max_H \mathcal{F}_K(q(X)q(\Theta), H^{(i)}) \tag{7}$$

As an alternative, hierarchical priors may be considered. In this approach, one may attach priors to the hyperparameters as well, that are governed by hyper-hyperparameters, and so on. Thus, one may consider marginalizing w.r.t to the parameters instead of maximizing. This approach is used in (Kenny, 2010) where a vague Gamma (hyper)-prior is attached to the precision of the Gaussian prior, resulting in an overall student-t prior distribution of the speaker factor. The experimental results of the 2010 speaker verification competition of NIST showed that the inclusion of this additional level of hierarchy increases significantly the verification accuracy.

4. A Variational Bayes approach to speaker diarization using supervectors

In this section, we examine in detail a VB approach to SD that utilizes supervectors in order to represent speech segments. Supervectors are high-dimensional vectors that are formed by concatenating the mean values of a GMM. The GMM is MAP-adapted from a Universal Background Model (UBM), where only the means are allowed to be adapted (see Appendix). Each supervector is then projected onto a space of lower dimensionality and VB inference in

adopted to estimate the number of speakers and the assignment of segments to speakers. VB methods that do not make use of the supervector representation can be found in (Valente, 2005),

4.1 Supervectors and modeling assumptions

As explained in the Appendix, supervectors are high-dimensional vectors that are capable of capturing speaker characteristics in great detail, and are applied to speaker verification and recently in diarization, too. A main assumption used in the proposed method is that a supervector M can be described by a mid-dimensional vector w, as follows

$$M \approx M_0 + Vw \tag{8}$$

where M_0 the center of the acoustic space (i.e. the supervector of the UBM) and V a low rank (say p) rectangular matrix. The columns of V are the eigenvectors and have been extracted off-line. Furthermore, the columns of V are properly scaled with the corresponding eigenvalues so that $w \sim \mathcal{N}(0, I_p)$. Finally, let Σ be the diagonal covariance matrix of M_0 (see (Kenny et al., 2005) for a detailed derivation).

Let us assume (i) that a segmentation of the stream Y into segments has been applied. A uniform segmentation of $1s$ duration is proposed in (Kenny et al., 2005), however, speaker change detection techniques may be applied as well. The segmented MFCC stream is denoted by $Y = \{y_m\}_{m=1}^{M}$. For a given number of speaker K, an K-dimensional indicator vector i_m is used to indicate the speaker it belongs to, that is $i_{mk} = 1$, if and only if y_m belongs to the kth speaker. The collection of these vectors is denoted by $\mathcal{I} = \{i_m\}_{m=1}^{M}$. Moreover, the parameter vector of the kth speaker is denoted by w_k and their collection as $W = \{w_k\}_{k=1}^{K}$.

We further assume (ii) that an upper bound of the number of speakers (say K_{max}) is given and that the mixing coefficients $\pi = \{\pi_k\}_{k=1}^{K}$ (i.e. the prior probabilities of each speaker) can be estimated by maximizing the marginal likelihood

$$\int p(Y, W, \mathcal{I}|\pi) dW d\mathcal{I} \tag{9}$$

w.r.t. π. This technique, known as Maximum Likelihood II (ML-II) clearly diverges from the Bayesian framework. A fully-Bayesian approach attaches priors (e.g. Dirichlet) to $\{\pi_k\}_{k=1}^{K}$ and integrates out these parameters, too, instead of maximizing w.r.t. them. However, this technique enables us to estimate K without resorting to comparison between the marginal likelihood of several K, which can be time consuming when dealing with a large range of candidate number of speakers. On the contrary, by using this technique, we can estimate K simply by counting the number of mixture coefficients assigned non-zero values by ML-II, (Corduneanu & Bishop, 2001).

Finally, we assume (iii) that the alignment of frames with GMM-level mixture components is given. This assumption uses the final E-step of the EM algorithm as an estimate of the missing data (i.e. the component indicators). Using this assumption, we not only have to deal with the a single set of missing data, i.e. $\mathcal{I} = \{i_m\}_{m=1}^{M}$, but we are able to represent segments with *sufficient statistics* and utilize closed-form expressions to calculate the desired statistical quantities. This is due to the fact that the complete-data likelihood of a GMM belong to an exponential family, while the incomplete-data likelihood does not.

4.2 Working with the complete-data

To stress the benefits from the third assumption, let us derive some useful formulae that will be used, namely the likelihood, the posterior and the marginal likelihood of a *single* GMM that is represented by w. Let $y_u = \{y^t\}_{t=1,2,...}$ be the MFCC coefficients of a segments. We parametrize the (centralized) statistics of each segment as

$$N_c = \sum_t \gamma^t(c) \tag{10}$$

$$\tilde{F}_c = \sum_t \gamma^t(c) \left(y^t - \mu_c^0\right) \tag{11}$$

and

$$\tilde{S}_c = diag\left(\sum_t \gamma^t(c) \left(y^t - \mu_c^0\right) \left(y^t - \mu_c^0\right)^T\right) \tag{12}$$

where $\gamma^t(c)$ the posterior probability that y^t belongs to the cth component, given by the MAP-EM algorithm. This is our estimate of the missing data, that is already in-hand from the MAP-EM algorithm. For notational compactness, let us define $\mathbf{N}$ the $Cd \times Cd$ diagonal matrix, whose C diagonal block are defined as $\{N_c I_d\}_{c=1}^C$. Let also $\tilde{F}$ a Cd dimensional vector (i.e. a centralized supervector) by concatenating all $\tilde{F}_c$ and finally, let $\tilde{S}$ be the $Cd \times Cd$ diagonal matrix, whose C diagonal block are $\{\tilde{S}_c\}_{c=1}^C$.

To calculate the complete-data *likelihood* of a model with fixed parameters w given y_u, the following closed form expressions can be utilized

$$\log p(z_u|w) = G + H(w) \tag{13}$$

where

$$G = -\frac{1}{2}\text{tr}\left(\Sigma^{-1}\tilde{S}\right) - \sum_{c=1}^{C} N_c \log|2\pi\Sigma_c|^{1/2} \tag{14}$$

and

$$H(w) = w^T V^T \Sigma^{-1}\left(\tilde{F} - \frac{1}{2}\mathbf{N}Vw\right) \tag{15}$$

and $z_u = (y_u, \gamma_u)$ the (estimated) complete data.

The *posterior* distribution of w given z_u is also Gaussian $w \sim \mathcal{N}(\tilde{w}, \Lambda^{-1})$, where

$$\tilde{w} = \Lambda^{-1} V^T \Sigma^{-1} \tilde{F} \tag{16}$$

and

$$\Lambda = I_p + V^T \Sigma^{-1} \mathbf{N} V \tag{17}$$

the precision matrix of the posterior. Recall that $w \sim \mathcal{N}(0, I_p)$ a priori.

Finally, the *marginal* likelihood $p(z_u|S=1)$ is given by the following formula

$$\log p(z_u|S=1) = \log \int p(z_u|w, S=1)p(w)dw = G - \frac{1}{2}|\Lambda| + \frac{1}{2}\tilde{F}^T\Sigma^{-1}V\Lambda^{-1}V^T\Sigma^{-1}\tilde{F} \tag{18}$$

The existence of the above closed-form expressions is a consequence of using the (estimated) complete-data likelihood instead of the incomplete-data likelihood.

4.3 The VB algorithm

In order to solve the intractable problem of estimating $\mathcal{I}$ and S, a VB can been developed. Assume again that the variational posterior that can be factorized as $Q(Y,\mathcal{I}) = Q(Y)Q(\mathcal{I})$. Note though that in this setting, all the posteriors are conditional on (i) the complete-data $\{z_m\}_{m=1}^M$ and (ii) on a point-estimate of π. To update this estimate a further step should be added to the general VB-EM iteration, which is the maximization of the marginal likelihood w.r.t π. We initialize our variables by setting K equal to the maximum number of speaker K_{max}, and by setting π as uniform, i.e. $\pi_m = \frac{1}{K}, m = 1,\ldots,M$.

The E-step is responsible for estimating the assignment $\mathcal{I}$ given the current posterior distribution of the parameters $\{w_k\}_{k=1}^K$ and the current point-estimate π. Note that due to the conditioning on π, the factorization $Q(\mathcal{I}) = \prod_{m=1}^M Q(i_m)$ ($\{i_m\}_{m=1}^M$ are conditionally i.i.d.). Using the general update rule in (3) and after some matrix algebra, we end-up with

$$\text{VB-E step: } Q(i_m) = \prod_{k=1}^K q_{mk}^{i_{mk}}, \text{ where } q_{mk} = \frac{\tilde{q}_{mk}}{\sum_{k'=1}^K \tilde{q}_{mk'}} \tag{19}$$

and

$$\tilde{q}_{ms} = \pi_k p(z_m|\tilde{w}_k) \exp\left(-\frac{1}{2}\text{tr}\left(V^T \mathbf{N}_m \Sigma^{-1} V \Lambda_k^{-1}\right)\right) \tag{20}$$

where $p(z_m|\tilde{w}_k)$ and Λ_k are given in (13) and (17), respectively. Both quantities are estimated during the M-step of the previous iteration. Moreover, note as $\text{tr}\left(\Lambda_k^{-1}\right) \to 0$, i.e. no uncertainty is assumed regarding the estimates, the E-step degenerates to the corresponding step of the MAP-EM.

Similarly, the VB-M step is given according to the general rule in (4). After some matrix algebra we obtain

$$\text{VB-M step: } Q(w_k) \sim \mathcal{N}\left(\tilde{w}_k, \Lambda_k^{-1}\right) \tag{21}$$

i.e. will be a normal distribution with mean $\tilde{w}_k$ and precision Λ_k given in (16) and (17), respectively. Note again that the M-step of the MAP-EM is recovered by letting $\text{tr}\left(\Lambda_k^{-1}\right) \to 0$. Finally, the additional step for re-estimating π is derived by maximizing the marginal likelihood w.r.t π. By rejecting irrelevant terms, the maximization problems becomes the following

$$\hat{\pi} = \arg\max_{\pi} \sum_{m=1}^M \sum_{k=1}^K q_{mk} \log \pi_k, \text{ subject to } \sum_{k=1}^K \pi_k = 1 \tag{22}$$

which yields

$$\pi \text{ update step: } \hat{\pi}_k = \frac{1}{M} \sum_{m=1}^M q_{mk} \tag{23}$$

By iteratively applying (19), (21) and (23) the algorithm converges to a maximum. After convergence is reached, the assignment of segments to speakers $\mathcal{I}$ and consequently the number of speakers K are estimated from (19) by simply assigning the mth segments to the speaker that maximizes $Q(i_m)$.

4.4 Experiments

So far, the proposed system has been tested only against telephone conversation datasets. This setting differs from the usual diarization systems, since we a priori know the number of speakers (i.e. $K = 2$). Therefore, the strength of the proposed VB-system as a model selection tool cannot be assessed from this series of experiments. However, the results show a drastic reduction in terms of Diarization Error Rate (DER,%). In Table 1, the DER on NIST 2008 SRE Summed Channel Test Data made by the VB-system are presented, for several front-end features. Details about the features can be found in (Kenny et al., 2010). The most

	configuration	*mean* DER(%)	σ(%)
	BUT features		
1	VB without Viterbi	9.1	11.9
2	VB with Viterbi	4.5	8.5
3	VB with Viterbi and 2nd pass	3.8	7.6
	CRIM features		
4	VB with 2nd pass, no Viterbi	3.3	7.8
	Raw cepstral features		
5	VB with 2nd pass, no Viterbi	2.2	5.8
6	VB with 2nd pass, no Viterbi	1.9	5.6

Table 1. DER (%) NIST 2008 SRE Summed Channel Test Data using the VB-system. The standard deviation of the Diarization Errors is denoted by σ.

successive front-end configuration includes 20 static-only MFCC, a 1024-component UBM and a gender-independent factor analysis model with 300 eigenvoices. The 2nd pass means that the speaker change points found by Viterbi resegmentation were used to initialize a second run of Variational Bayes and this was followed by another Viterbi resegmentation.
In Table 1, the best performance of the VB-system is compared to (i) a baseline diarization system (i.e. speaker change detector, BIC-based AHC with single Gaussians and Viterbi resegmentation) augmented by a soft-clustering postprocessing stage, and (ii) a streaming system that operates on speaker factors and was introduced in (Castaldo et al., 2008) as a stream-based approach to performs online diarization. The conversation is seen as a stream of fixed-duration time slices and the system operates in a causal fashion. Speakers detected in the current slice are compared with previously detected speakers to determine if a new speaker has been detected or previous models should be updated. Further details about the implementation may be found in (Kenny et al., 2010).

System	*mean DER(%)*	σ(%)
Baseline with soft-clustering	3.5	8.0
Streaming with Viterbi	4.6	8.8
VB with raw cepstra, Viterbi and 2nd pass	1.0	3.5

Table 2. Best results obtained on the NIST 2008 SRE Summed Channel Telephone Data using the baseline, the streaming and the VB systems.

5. An HMM-based approach using hierarchical dirichlet processes

In this chapter, we present a recent SD approach that is based on Bayesian nonparametric modeling, (Fox et al., 2009). This approach utilizes the HMM framework to model the inter-speaker dynamics, mixture models for the emission probabilities and (averages of) MFCC as front-end-features. Its main contribution is the use of infinite models on both of the HMM levels, i.e. on the multimodal emission probabilities and on the states and the transitions between them.

5.1 General about infinite models

The use of infinite models is a natural way to overcome the issue of how to determine a priori the order of a model. First, consider the problem of determining the order of the GMM that should be used to model the distribution of a speaker. A fully Bayesian modeling should consider the order of the model as a random variable, and treat it in the same way it treats the rest of the parameters; the order should be integrated out, too, just like the weights, the means and the covariance matrices. On the HMM-level, a classical approach to determine the number of states K is to apply Viterbi, Baum-Welch or VB-EM type of learning for each of the candidate K by conditioning on K (i.e. on the hypothesis), and select the order that maximizes the evidence (or an approximation) of the model. The Evolving-HMM and the VB approach of (Valente, 2005) are typical examples of this framework. However, such exhaustive search solutions may lack of efficiency, especially in cases where the hypothesis space is quite large (e.g. Broadcast News). A more flexible solution is offered by infinite HMM, where the number of states are not specified a priori, but is rather inferred in a more data-driven way.

5.2 Infinite mixture models and the Dirichlet processes

We begin the analysis by describing the Dirichlet process (DP), which is the building block in most of the infinite models, (Ferguson, 1973). The DP can be considered as a infinite extension of the Dirichlet distribution. In the same way the Gaussian process can be utilized in Bayesian inference as a prior (i.e a measure) on functions, the DP can be used as a measure on measures. Moreover, much like the derivation of the familiar Gaussian process from the Gaussian distribution, the DP may be explicitly derived from the Dirichlet distribution by letting its order go to infinity.

Let us assume that $\beta = \{\beta_k\}_{k=1}^K$ follows a symmetric Dirichlet distribution of order K and strength α_0, i.e.

$$\beta|\alpha_0 \sim \mathcal{D}ir(\underbrace{\alpha_0/K, \ldots, \alpha_0/K}_{K \text{ components}}) \tag{24}$$

where $0 < \beta_k < 1, k = 1, \ldots, K$ and $\sum_{k=1}^K \beta_k = 1$. Suppose we aim to construct a generative model for GMMs. Then, β can be used as the weights of the model. We also need an appropriate *base* measure G_0. In the DP-GMM case, G_0 is the prior distribution of the components, e.g. a Normal-Inverse Wishart distribution if conjugacy is desired. By sampling the measure G_0 K times, i.e. $\theta_k \sim G_0, k = 1, \ldots, K$ we get a set of K *atoms* that can be associated with $\{\beta_k\}_{k=1}^K$. The distribution of θ given $\{\beta_k, \theta_k\}_{k=1}^K$ can expressed as $\theta|G^k \sim G^k$ where

$$G^K = \sum_{k=1}^K \beta_k \delta_{\theta_k} \tag{25}$$

where $\delta_{\theta_k} = \delta(\cdot, \theta_k)$. The distribution G^K can now be used in order to generate random samples from G^K. One should first sample $\theta|G^K \sim G^K$ and then sample $y|\theta \sim F(\theta)$, where $F(\cdot)$ is the Gaussian distribution.
Suppose now that we let $K \to \infty$. Then, $\beta|\alpha_0$ follows a DP with concentration parameter α_0. The random draw from the DP becomes an infinite mixture, i.e.

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{26}$$

We say that G follows a DP and we denote it by

$$G \sim \text{DP}(\alpha_0, G_0) \tag{27}$$

Let us consider N samples from G, denoted by $\{\phi_n\}_{n=1}^N$, $\phi_n|G \sim G$. What prevents K from going to infinity as $N \to \infty$ is a fundamental property of the Dirichlet distribution $\mathcal{D}ir(\{g_k\}_{k=1}^K)$. Starting from $g_k = 1, k = 1, \ldots, K$ and letting $g_k \to 0$, the probability mass is being increasingly concentrated on areas close to the K *vetrices* of the $(K-1)$-simplex. Hence, even if $N \to \infty$, G remains discrete and the cardinality of the set finite.
The posterior of G, i.e. conditioned to a set $\{\phi_n\}_{n=1}^N$ is a DP, parametrized as follows

$$G \sim \text{DP}\left(\alpha_0 + N, \frac{1}{\alpha_0 + N}\left[\alpha_0 G_0 + \sum_{n=1}^{N} \delta_{\phi_n}\right]\right) \tag{28}$$

or equivalently

$$G \sim \text{DP}\left(\alpha_0 + N, \frac{1}{\alpha_0 + N}\left[\alpha_0 G_0 + \sum_{k=1}^{K} N_k \delta_{\theta_k}\right]\right) \tag{29}$$

where $N_k = \sum_{n=1}^{N} \delta(\phi_n, \theta_k)$ and $\sum_{k=1}^{K} N_k = N$.
It order to create samples from the DP, we may proceed as follows

$$\phi_{N+1}|\{\phi_n\}_{n=1}^N \sim \frac{1}{\alpha_0 + N}\left(\alpha_0 G_0 + \sum_{n=1}^{N} \delta_{\phi_n}\right) \tag{30}$$

i.e. there is no need to refer to G. What (30) shows is that as N grows, the probability of getting previously unseen samples decreases linearly. Furthermore, the probability of the new sample to be equal to θ_k is equal to $(\alpha_0 + N)^{-1} N_k$. Finally, high values of α_0 corresponds to high rates of generating unseen atoms.
Given α_0, the prior of the number of distinct atoms K after N samples is given by

$$p(K|\alpha_0, N) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} s(N, K) \alpha_0^K \tag{31}$$

where $s(N, K)$ are unsigned Stirling numbers of the first kind.
A intuitive and constructive interpretation of β is the *stick-breaking* process, (Sethuraman, 1994). For the finite case, we saw that β follows the Dirichlet distribution. In order to create samples of β for the infinite case, however, the following sampling scheme is useful. Considering a stick of unitary length. For $k = 1, 2, \ldots$,

$$u_k|\alpha_0 \sim \mathcal{B}eta(1, \alpha_0) \tag{32}$$

$$\beta_k = u_k \left(1 - \sum_{k'=1}^{k-1} \beta_{k'}\right) = u_k \prod_{k'=1}^{k-1} (1 - u_{k'}) \tag{33}$$

Therefore, u_k is distributed as $\mathcal{B}eta(1, \alpha_0)$ and covers a fraction of u_k of the remaining stick. Hence, the overall length that covers is equal to β_k, given by (33). Note that a usual notation from the stick-breaking weights is $\beta \sim \text{GEM}(\alpha_0)$, where GEM stands for Griffiths, Engen and McCloskey. The generative model is depicted in Fig. 2(a).

5.3 Infinite Hidden Markov Models and the hierarchical DP

Let us now examine how can we apply similar ideas to a dynamic network, namely the (time-independent) HMM. An HMM can be considered as a collection of GMMs, that differ only on their *weights* which correspond to the *rows* of the transition matrix A. Each row $A_{k\cdot} = [A_{k1}, \ldots, A_{kK}]$, $k = 1, \ldots, K$ is the conditional probability of $x^{t+1} = l$, $l = 1, \ldots, K$ given $x^t = k$. Moreover, the initial probabilities $a = [a_1, \ldots, a_K]$ may also be treated in a similar way, by defining the non-emitting zero state. This allows us to include all the transition parameters in a unique matrix, defined as the augmented transition matrix $A^+ = [a^T, A^T]^T$.
In the finite-state case, a standard Bayesian strategy is to place a common prior on each line of A, e.g.

$$p(A_{k\cdot}|\gamma) = \mathcal{D}ir(\gamma/K, \ldots, \gamma/K) \tag{34}$$

Two are the drawbacks of this approach. The first is that the state-persistence that several dynamic systems exhibit is not captured explicitly in this prior. As we show next, this can be solve rather easily, by adding an extra hyperparameter to the diagonal of A that is capable of biasing the dynamics towards self-transition. A further and more severe in our case drawback is that such a prior cannot be extended to the infinite case. This is because the tying between the weights $A_{k\cdot} = [A_{k1}, \ldots, A_{kK}]$, $k = 1, \ldots, K$ that is offered by placing a common prior is weak when $K \to \infty$. What this prior implies is that each $A_{k\cdot}$ should simply be a independent draw from a DP having a common concentration parameter γ and a common continuous base measure H. Hence, the set of atoms between every pair of draws would be disjoint, leading to no sparse solutions at all. As proposed in (Teh et al., 2006), what is required to tackle this

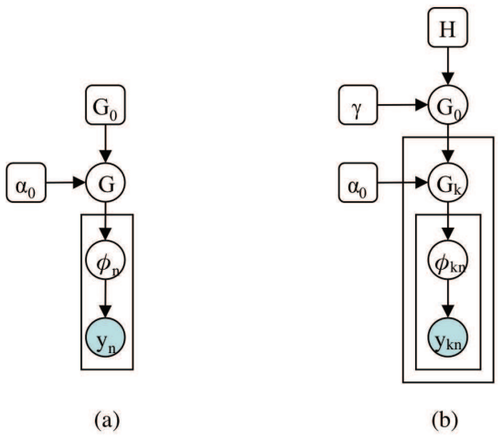

Fig. 2. Plate notations of the DP-mixtures. (a) The original DP-mixture model, (b) The Hierarchical DP-mixture model

problem is to add another level in the hierarchy. On the uppermost level, a single draw G_0 from $DP(\gamma, H)$ is generated, i.e.

$$G_0|\gamma, H \sim DP(\gamma, H) \tag{35}$$

This draw is then used to parametrize the DP prior of each of the states, i.e.

$$G_k|\alpha_0, G_0 \sim DP(\alpha_0, G_0),\ k = 1, \ldots, \infty \tag{36}$$

The generative model is depicted in Fig. 2(b). Contrary to the previous approach, the base measure of G_k (denote by G_0) is not only common to all states, but is moreover discrete, since

$$G_0|\gamma, H = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{37}$$

Hence, each G_k would be a (weighted) collection of the same set of atoms. Moreover, not only the set is the same, but identically weighted by $\beta = \{\beta_k\}_{k=1}^{\infty}$. Using the stick-breaking construction, each row of the transition matrix is distributed as follows. For $k = 1, 2, \ldots$ and $k' = 1, 2, \ldots$

$$u_{kk'}|\alpha_0 \sim \mathcal{B}eta\left(\alpha_0\beta_{k'}, \alpha_0\left(1 - \sum_{l=1}^{k'} \beta_l\right)\right) \tag{38}$$

$$A_{kk'} = u_{kk'} \prod_{l=1}^{k'-1} (1 - u_{kl}) \tag{39}$$

The expected values of each $A_{k\cdot}$ will be equal to β_k. Moreover, the concentration α_0 now controls both the state-connectivity and the similarity between each $A_{k\cdot}$. High values of α_0 means that most of the samples will be generated directly from G_0, which increases the state connectivity and decreases the variability between $\{A_{k\cdot}\}_{k=1}^{K}$. Contrarily, for low values of α_0 the HMM may exhibit sparse state-connectivity, i.e. each state may be accessible only via a subset of the other states.

5.4 Hierarchical DP HMM with DP mixture models as emission probabilities

Let us recapitulate the above modeling. We showed that the Hierarchical DP is a natural extension of the original DP, that is suitable in cases where the overall model is decomposed to a collection of submodels that share some certain properties. HMMs are such models, since they can be considered as collections of conditional mixtures, where the conditioning is w.r.t. the current state. We emphasize that these mixtures should not be confused with the possibility of modeling the emissions probabilities with mixture models. The emission probabilities are governed completely by the base measure $H(\cdot)$. If we desire to include finite mixtures (e.g. GMMs) then $H(\cdot)$ should be the Dirichlet-Normal-Inverse Wishart prior distribution, if conjugacy is desired. The tth observation y^t will then follow a GMM distribution, $y^t|\theta_{x^t} \sim F(\theta_{x^t})$. Thus, each atom θ_k will be a parametrization of a GMM, capable of describing the multimodal distribution of a speaker.

For describing the distribution of a speaker, the use of DP-mixture models may be considered as well. This means that both the HMM and its emissions may be considered as infinite (i.e. nonparametric), which is the method proposed in (Fox et al., 2009). However, in order to avoid fast transitions between states, a bias towards self-transitions is adopted, that allows to distinguish between the underlying HDP-HMM states and the within-speaker multimodal

emissions. Moreover, non-overlapping 250*ms* frames are used as front end features while a minimum duration of 500*ms* is imposed on speaker segments. The resulting model, termed as the *Sticky*-HDP-HMM produced state-of-the-art results even without any prior tuning. In fully Bayesian approaches, tuning is related to the hyperparameters of the uppermost layer. We also emphasize that the use of infinite models is SD has previously been proposed in (Valente, 2006). It uses a DP-mixture model for the emissions and an Infinite-HMM for modeling the transition dynamics. However, the HMM used was degenerated (all rows of A are assumed to be equal) making the hierarchical DP unnecessary. A VB algorithm was proposed, based on the mean-field approximation, while a slight improvement was reported over the baseline VB with finite mixtures.

5.5 Inference

Methods of inference of the Sticky-HDP-HMM is out of the scope of this review. The interested reader is encouraged to examine the inferential procedures given in (Teh et al., 2006) and (Fox et al., 2009).

In general, to infer such models, the most usual way is the family of Markov Chain Monte Carlo (MCMC) methods. Like any sampling method, MCMC aims to estimate any desired quantity by sample averages, generated according a proper measure. In cases where all of the distributions are conjugate to their priors, Gibbs sampling is usually a sufficient and easy to implement MCMC method. It proceeds with sampling each random variable, conditioned on all the others, which are set to their current values. The Gibbs sampler is not a unique technique in the models described above. This is because there are alternative generative models by which the same process can be defined. Several Gibbs samplers have been proposed, that vary according to their mixing rates and their implementation effort that is required. A detailed implementation of these such samplers, along with a comparison between them can be found in (Teh et al., 2006). Other approaches, that are better suited the HMM framework are presented in (Fox et al., 2009). Finally, we mention the possibility of applying variational inference to infinite models. Such approaches are analyzed in (Blei & Jordan, 2005) and (Valente, 2006) and can be much faster that MCMC.

5.6 Experiments

The experiments of the Sticky-HDP-HMM presented in (Fox et al., 2009) are based on the NIST-2007 meeting data and are being compared to (i) the non-sticky-HDP-HMM and to (ii) the ICSI diarization system, (Wooters & Huijbregts, 2008). The latter system is based on AHC and was the winner of the competition, scoring a 18.37% DER. It uses ML-GMMs to model the emission probabilities, a penalty free BIC-like approach and a Viterbi algorithm after each cluster merging. The comparison between the two HDP systems is presented in Table 3. The number in the parentheses is the performance when running the 16^{th} meeting for 50,000

Overall DERs (%)	Min Hamming	Max Likelihood	2-Best	5-Best
Non-Sticky HDP-HMM	23.91	25.91	23.67	21.06
Sticky HDP-HMM	19.01 (17.84)	19.37	16.97	14.61

Table 3. Best results obtained on the NIST-2007 Meeting Data using the Sticky and the Non-Sticky HDP-HMM.

Gibbs iterations, instead of the fixed number of 10,000 iterations. The results clearly show the usefulness of the state persistence parameter in avoiding the unrealistic fast transitions between speakers that is translated to an approximate 20% relative improvement in DER.

Compared to the ICSI system, the Sticky-HDP-HMM performed slightly worst, if we consider the setting with 10.000 iterations. We should note though that no tuning has been applied, i.e. the priors on the hyperparameters are very vague, and are therefore placing significant prior mass over areas that are unrealistic for the specific application field. Hence, by assuming a proper tuning of the uppermost hyperperameters, a further increase in the accuracy should normally be expected.
Note finally, that due to the fully Bayesian paradigm, several alternative state-sequences may be sampled from the posterior. As Table 3 shows, if the best per-meeting DER for the five most likely samples is considered, our overall DER drops to 14.61%. Finally, the possibility of providing multiple state-sequences, along with their posterior probability mass, is a desirable property when applying fusion techniques. In such cases, the relative uncertainty of the decisions made by each information stream should also be assessed in order to fuse the streams in a fully probabilistic manner.

6. Conclusions and further research directions

In this chapter, we presented a introduction to some of the recent methods that have been proposed in SD. We restricted ourselves to some novel fully-Bayesian approaches, that are based on (i) finite mixtures with Variational Bayes inference methods, and (ii) nonparametric (i.e. infinite) Bayesian approaches. These methods are applicable to numerous problems that deal with clustered data and are gaining increasing attention in several fields. We analyzed some of the theoretical advantages over non- or semi-Bayesian approaches and their strength and flexibility in learning the clustered structures of the data.
Bayesian nonparametrics may be used to tackle several other tasks in speaker and audio problems, as well. For example, speaker verification is another major task that can be treated as a model selection problem (that is one versus two speakers), and the effectiveness of fully-Bayesian approaches has recently been proven, (Kenny, 2010). Furthermore, SVM-based verification is a field where Bayesian approaches can be examined. A severe problem with SVMs is that their soft-outputs cannot be regarded as probabilistic. On the contrary, relevance vector machines (RVM) are fully-probabilistic analogues to SVMs and as such they can be used as an alternative discriminative framework, (Tipping, 2001). Speaker separation from multiple (or single) microphones is another related task to SD. A Bayesian nonparametric model, termed as infinite factorial HMM has been used to separate the speakers and infer their number, (Van Gael et al., 2009). Such approaches can be used in SD as well, in order to detect and identify overlapping speakers. Finally, several inference methods can be tested in speaker technologies, such as the Annealing Importance Sampling (Neal, 1998) and Expectation-Propagation (Minka, 2001) that produce state-of-the-art results in many other fields.

7. Appendix: Super- and i-vectors feature spaces

We review here some of the new feature spaces that are used in most of the contemporary speaker verification systems and recently in several SD systems as well. These features are derived by (i) adapting a UBM with the observation vectors of a speech segment (using the standard EM-MAP of (Reynolds et al., 2000)) and (ii) mapping the high-dimensional concatenated mean vector (or *supervector*) to a mid-dimensional subspace, resulting in the identity vector, or simply the *i-vector*. The transformation rule is derived offline, using enrollment data, and aims to reduce the dimensionality of the new feature space, while

discarding those directions that do not carry speaker discriminant information.
The major advantage of the new feature space is the mapping of variable-length utterances onto a space of fixed dimensionality, through a well-tested statistical intermediate description (i.e. the UBM-based adaptation scheme). Using the i-vector representation, several kernel-based and other general purpose algorithms can be applied in order to perform identification, verification, and clustering. Finally, the i-vectors of a speaker have a rather Gaussian distribution since they represent mean values, projected onto a lower dimensional basis and they take values on $\Re^p$. Hence, several algorithms that have been developed assuming Euclidean spaces (i.e. of constant metric tensor) can be applied without much adaptation. This is in contrast to representations that lie on spaces where the natural statistical divergences (e.g. KL, Hellinger) have complex expressions that are far from being (squared) Euclidean distances, such as those that include weights or covariance matrices.

MAP-estimate based on UBM and the supervector representation

As discussed in section 2, a typical preprocessor applies MFCC extraction, delta-feature calculation and voice activity detection. When performing speaker verification, normalization methods such as mean and variance normalization, RASTA filtering and feature warping are essential in order to compensate for the channel-effects, (Kinnunen & B, 2010).
An effective statistical representation of the stream $Y = \{y^t\}_{t=1}^T$ of front-end features is a Gaussian Mixture Model (GMM). The model, however, is not trained from scratch. Instead of a Maximum Likelihood (ML) estimate, the observations are used to adapt a well-trained model (Universal Background Model, UBM) with parameters $\lambda_{ubm} = \{\pi_c^0, \mu_c^0, \Sigma_c^0\}_{c=1}^C$ that denote weights, means and (diagonal) covariance matrices, respectively. The UBM is a GMM that is trained offline with the standard ML-EM algorithm, using hours of speech data and a huge number of speakers. The final estimate $\tilde{\lambda}_Y$ of the p.d.f. of Y is the Maximum A Posteriori (MAP) estimate of λ_Y, and is calculated by a MAP-Expectation-Maximization (MAP-EM) algorithm. Moreover, only the mean-values are allowed to be adapted, which implies that the mean values $\{\tilde{m}_c\}_{c=1}^C$ are sufficient to represent the model $\tilde{\lambda}_Y$ for a fixed UBM.
The E-step of the ith iteration is carried as

$$P(c|y^t, \tilde{\lambda}_Y^{(i-1)}) = \frac{\pi_c^0 p(y^t|\tilde{\mu}_c^{(i-1)}, \Sigma_c^0)}{\sum_{c=1}^C \pi_c^0 p(y^t|\tilde{\mu}_c^{(i-1)}, \Sigma_c^0)} \tag{40}$$

followed by the corresponding M-step

$$\tilde{\mu}_c^{(i)} = \alpha_c^{(i)} \bar{y}_c + (1 - \alpha_c^{(i)}) \mu_c^0 \tag{41}$$

where

$$\alpha_c^{(i)} = \frac{n_c^{(i)}}{n_c^{(i)} + r} \tag{42}$$

$$n_c^{(i)} = \sum_{t=1}^T P(c|y^t, \tilde{\lambda}_Y^{(i-1)}) \tag{43}$$

and

$$\bar{y}_c = \frac{1}{n_c^{(i)}} \sum_{t=1}^T P(c|y^t, \tilde{\lambda}_Y^{(i-1)}) y^t \tag{44}$$

The above expressions reveal that λ_{ubm} and r are completely specifying the *prior* of $\{\mu_c\}_{c=1}^{C}$. Its density is as follows

$$\mu_c|r,\Sigma_c \sim \mathcal{N}\left(\mu_c^0, \frac{1}{r}\Sigma_c^0\right) \tag{45}$$

where $\mathcal{N}(\mu,\Sigma)$ denotes the normal p.d.f., with mean and covariance matrix μ and Σ respectively. The parameter r corresponds to the *strength* of the prior of $\{\mu_c\}_{c=1}^{C}$, i.e. the equivalent number of virtual observations that are backing the initial estimate μ_c^0.
Apart from the increase in the robustness of the estimate of λ_Y, a further severe benefit from using a UBM is the common *ordering* that it establishes to the C open areas of the observations' space. Consider the MAP estimates $\tilde{\lambda}_{Y_a}$ and $\tilde{\lambda}_{Y_b}$ given Y_a and Y_b respectively. Due to their common initialization by λ_{ubm}, $\tilde{\lambda}_{Y_a}$ and $\tilde{\lambda}_{Y_b}$ are directly comparable, in the sense that their corresponding entries carry information about the same a priori area of the observations' space, apart from the dimension. Such a correspondence cannot be achieved when the models are trained using ML-EM algorithm, making several fast scoring methods and dimensionality reduction techniques inapplicable. The concatenated vector $M_Y \in \Re^{Cd}$ of the means $\{\tilde{\mu}_c\}_{c=1}^{C}$ is termed *supervector* and can be considered as a novel fixed-size way for representing Y.

Likelihood ratios in verification and clustering

A standard way to score a new set of observation against a model $\tilde{\lambda}_s$ is based on the normalized log-likelihood ratio between the $\tilde{\lambda}_s$ and λ_{ubm}, i.e.

$$NLLR(\tilde{\lambda}_s|Y,\lambda_{ubm}) = \frac{1}{T}\sum_{t=1}^{T}\log\frac{p(y^t|\tilde{\lambda}_s)}{p(y^t|\lambda_{ubm})} \tag{46}$$

The coupling between $\tilde{\lambda}_s$ and the UBM increases drastically the robustness of the ratio, and allows fast scoring methods to be applied.
The NLLR can be deployed in order to apply both verification and clustering. In verification, NLLR is normalized properly according to a set of cohort speakers and then a simple threshold is applied to verify the claimed identity. Several score-level normalization methods have been proposed (e.g. z-norm, t-norm, s-norm) and are aiming to compensate the speaker and channel dependent behavior of the statistic NLLR.
In most step-by-step SD approaches, UBM-based models are used only after a first clustering pass with single-Gaussian models. The clusters that are created are then used to initialize further iterations of UBM-based hierarchical clustering. To define a similarity measure between two clusters Y_a and Y_b, the Cross Likelihood Ratio (NCLR)

$$CLR(Y_a,Y_b) = NLLR(\tilde{\lambda}_a|Y_b,\lambda_{ubm}) + NLLR(\tilde{\lambda}_b|Y_a,\lambda_{ubm}) \tag{47}$$

and the Normalized Cross Likelihood Ratio (NCLR)

$$NCLR(Y_a,Y_b) = \frac{1}{T_a}\sum_{y^t\in Y_a}\log\frac{p(y^t|\tilde{\lambda}_b)}{p(y^t|\tilde{\lambda}_a)} + \frac{1}{T_b}\sum_{y^t\in Y_b}\log\frac{p(y^t|\tilde{\lambda}_a)}{p(y^t|\tilde{\lambda}_b)} \tag{48}$$

are both symmetric measures that have been applied successfully, (*see* (Le et al., 2007) for a comparison). However, a predefined threshold is required to decide whether a pair of clusters should be merged or not, (Zhu et al., 2005).

Kernels based on supervectors

One of the drawbacks of a likelihood ratio-based verification and clustering algorithms is their dependence on the data Y. This problem arises from the fact that the likelihood function of the *incomplete* data

$$p(y|\tilde{\lambda}) = \sum_{c=1}^{C} \pi_c p(y|\mu_c, \Sigma_c) \tag{49}$$

does not belong to an *exponential* family and, therefore, a sufficient statistic does not exist, (Wainwright & Jordan, 2008). On the contrary, the *complete-data likelihood*, i.e. the likelihood of $z^t = (x^t, y^t)$ where $X = \{x^t\}_{t=1}^{T}$ denotes the alignment of Y to components - belongs to the exponential family

$$p(x, y|\tilde{\lambda}) = \sum_{c=1}^{C} \delta(c, x)\pi_c p(y|\mu_c, \Sigma_c) \tag{50}$$

and therefore several closed-form expressions can be utilized. The obvious problem is that we do not know x^t. However, their MAP estimate $\tilde{x}^t$ of x^t is already in-hand, from the last E-step of the EM algorithm. This is the rationale for the use of similarity measures between utterances that are not based on likelihood-ratios. In (Campbell, Sturim & Reynolds, 2006), a KL divergence-like kernel that is proposed

$$K(\tilde{\lambda}_a, \tilde{\lambda}_b) = \sum_{c=1}^{C} \left(\sqrt{\pi_c}\Sigma_c^{-1/2}(\tilde{\mu}_c^a - \mu_c^0)\right)^T \left(\sqrt{\pi_c}\Sigma_c^{-1/2}(\tilde{\mu}_c^b - \mu_c^0)\right) \tag{51}$$

Such kernels implicitly make use of the complete-data likelihood, and the corresponding closed-form expressions. Once the kernel is defined, one may consider the use of Support Vector Machines (SVMs) to perform verification. During training, the separating hyperplane should be estimated, based on a labeled training set that consists of both positive and negative examples $\{\tilde{\lambda}_i, t_i\}_{i=1}^{N}$, where $t_i \in \{-1, +1\}$. During verification, a sparse subset Λ_s of these examples (i.e. the support vectors) $\{\tilde{\lambda}_i, t_i\}_{i\in\Lambda_s}$ along with their weights $\{\alpha_i\}_{i\in\Lambda_s}$ and the bias term b are needed to perform verification, according to $sgn(f(\tilde{\lambda}'))$, where

$$f(\tilde{\lambda}') = \sum_{i\in\Lambda_s} \alpha_i t_i K(\tilde{\lambda}', \tilde{\lambda}_i) + b \tag{52}$$

denotes the function that defines the hyperplane. Several other kernels and additional information regarding the SVM-based verification can be found in (Campbell, Campbell, Reynolds, Singer & Torres-Carrasquillo, 2006).

From supervectors to i-vectors

In practice, the dimensionality of supervectors is very large to handle (e.g. $dim(M_Y) = 77824$ for $(d, C) = (38, 2048)$). Therefore, it is a natural field for applying dimensionality reduction (DR) methods. A common method for DR is Principal Component Analysis (PCA). The eigenvectors having the highest corresponding eigenvalues are termed *eigenvoices*, inspired from the similar concept of eigenfaces in face recognition, (Turk & Pentland, 1991).

However, PCA is an *unsupervised* method, and as such, it does not take into account neither the clustered structure of the enrollment data nor the classification purpose of the DR. Linear Discriminant Analysis (LDA) is a popular *supervised* method for defining such bases and is

the one that is used to extract the i-vectors. The supervector M (of $\kappa = Cd$ dimensions) is assumed to be generated from the following equation

$$M = M_0 + Tw + e \tag{53}$$

where M_0 the supervector of the UBM, T a $(\kappa \times p)$-dimensional matrix (where $p \ll \kappa$, typically $p = 400$), w a p-dimensional vector having a standard normal distribution, i.e. $w \sim \mathcal{N}(0, I_p)$ and e the approximation error. The matrix T is called *total variability matrix* and its columns are forming the LDA-derived subspace with which M is expressed. The term total variability matrix stems from the fact that the labeling used in LDA treats each speaker recording (i.e. each set of utterances of a speaker from the same session) as a distinct class, (Dehak et al., 2011). This strategy is in contrast to a former one, that applies Joint-factor Analysis (JFA) to model separately between-speaker and within-speaker variability.
To calculate the i-vector w of an utterance u that consists of Y, assuming a UBM and a basis T, one should (i) adapt the UBM using the standard MAP-adaptation scheme, and (ii) use the centralized mean vectors to calculate the i-vector with the following formula

$$w = \left(I_p + T^T \Sigma_e^{-1} N_u T\right)^{-1} T^T \Sigma_e^{-1} F_u \tag{54}$$

In (54), F_u denotes the centralized supervector of the utterance, i.e. $F_u = M_u - M_0$, N_u a $\kappa \times \kappa$ diagonal matrix, whose K diagonal blocks are defined as $n_c I_d$ and n_c given in (43), and finally Σ_e a $\kappa \times \kappa$ diagonal covariance matrix, estimated during LDA, that models the expected variance of the approximation error e.
These vectors may be considered as lying on a feature space that is well suited to tasks like speaker verification, identification and diarization.

8. References

Ajmera, J. & Wooters, C. (2003). A robust speaker clustering algorithm, *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, pp. 411–416.

Amari, S. (1995). Information geometry of the EM and em algorithms for neural networks, *Neural Networks* 8: 1379–1408.

Anguera, X., Wooters, C. & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings, *IEEE TASLP* 15(7): 2011–2021.

Barras, C., Zhu, X., Meignier, S. & Gauvain, J. (2004). Improving Speaker Diarization, *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*.

Blei, D. M. & Jordan, M. I. (2005). Variational inference for Dirichlet process mixtures, *Bayesian Analysis* 1: 121–144.

Campbell, W., Campbell, J., Reynolds, D., Singer, E. & Torres-Carrasquillo, P. (2006). Support vector machines for speaker and language recognition, *Computer Speech and Language* 20(2-3): 210 – 229. Odyssey 2004: The speaker and Language Recognition Workshop - Odyssey-04.

Campbell, W. M., Sturim, D. E. & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification, *IEEE Signal Processing Letters* 13: 308–311.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P. & Vair, C. (2008). Stream-based speaker segmentation using speaker factors and eigenvoices, *Proc. IEEE ICASSP*, pp. 4133–4136.

Chen, S. & Gopalakrishnam, P. (1998). Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*.

Corduneanu, A. & Bishop, C. M. (2001). Variational bayesian model selection for mixture distributions, *Eighth Int'l Conf. Artificial Intelligence and Statistics*, pp. 27–34.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. & Ouellet, P. (2011). Front-end factor analysis for speaker verification, *Audio, Speech, and Language Processing, IEEE Transactions on* 19(4): 788 –798.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Ser. B* 39.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems, *Annals of Statistics* 1: 209–230.

Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2009). The Sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with Persistent States.

Fraley, C. & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, *Comput. J.* 41: 578–588.

Friedland, G., Vinyals, O., Huang, Y. & Müller, C. (2009). Prosodic and other long-term features for speaker diarization, *IEEE Transactions on Audio, Speech & Language Processing* 17(5): 985–993.

Gupta, V., Kenny, P., Ouellet, P., Boulianne, G. & Dumouchel, P. (2007). Combining Gaussianized/non-Gaussianized Features to Improve Speaker Diarization of Telephone Conversations, *IEEE Signal Processing Letters* 14(12): 1040–1043.

Hermansky, H., Hanson, B. A. & Wakita, H. (1985). Perceptually based linear predictive analysis of speech, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 509–512.

Hermansky, H., Morgan, N., Bayya, A. & Kohn, P. (1992). RASTA-PLP speech analysis technique, *Acoustics, Speech, and Signal Processing, IEEE International Conference on* 1: 121–124.

Ishizuka, K., Nakatani, T., Fujimoto, M. & Miyazaki, N. (2010). Noise robust voice activity detection based on periodic to aperiodic component ratio, *Speech Commun.* 52: 41–60.

Kenny, P. (2010). Bayesian speaker verification with heavy tailed priors, *Computer Speech and Language* . Odyssey 2010: The speaker and Language Recognition Workshop - Odyssey-10, Brno, Czech Republic.

Kenny, P., Boulianne, G. & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data., *IEEE Transactions on Speech and Audio Processing* 13(3): 345–354.

Kenny, P., Reynolds, D. & Castaldo, F. (2010). Diarization of telephone conversations using factor analysis, *Selected Topics in Signal Processing, IEEE Journal of* 4(6): 1059 –1070.

Kinnunen, T. & B, H. L. (2010). An overview of text-independent speaker recognition: from features to supervectors", speech communication.

Lapidot, I. (2003). SOM as likelihood estimator for speaker clustering, *in Proc. Eurospeech*.

Le, V.-B., Mella, O. & Fohr, D. (2007). Speaker Diarization using Normalized Cross Likelihood Ratio, *proceeding of INTERSPEECH 2007 INTERSPEECH 2007*, ISCA, Antwerp Belgium.

Mackay, D. J. C. (2003). Information theory, inference, and learning algorithms, *Cambridge University Press New York* .

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F. & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization, *Computer Speech and Language* 20: 303–330.

Minka, T. (2001). Expectation propagation for approximate bayesian inference, *Proceedings of the Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, Morgan Kaufmann, San Francisco, CA, pp. 362–36.

Neal, R. M. (1998). Annealed importance sampling, *STATISTICS AND COMPUTING* 11: 125–139.

Ning, H., Liu, M., Tang, H. & Huang, T. (2006). A Spectral Clustering Approach to Speaker Diarization, *Proceedings of the International Conference on Spoken Language Processing*.

Pardo, J., Anguera, X. & Wooters, C. (2007). Speaker diarization for multiple-distant-microphone meetings using several sources of information, *IEEE Trans. Comput.* 56: 1189–1224.

Pelecanos, J. & Sridharan, S. (2001). Feature warping for robust speaker verification, *ODYSSEY-2001*, pp. 213–218.

Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, Vol. 10, pp. 19–41.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica* 4: 639–650.

Snoussi, H. (2005). The geometry of prior selection, *Neurocomputing* 67: 214–244.

Stafylakis, T., Katsouros, V. & Carayannis, G. (2010a). The Segmental Bayesian Information Criterion and its applications to Speaker Diarization, *IEEE Selected topics in Signal Processing* pp. 857 – 866.

Stafylakis, T., Katsouros, V. & Carayannis, G. (2010b). Speaker clustering via the mean shift algorithm, *Odyssey 2010: The speaker and Language Recognition Workshop - Odyssey-10, Brno, Czech Republic*.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Hierarchical Dirichlet processes, *Journal of the American Statistical Association* 101(476): 1566–1581.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine, *Journal of Machine Learning Research* 1: 211–244.

Tranter, S. & Reynolds, D. (2006). An overview of automatic speaker diarization systems, *IEEE Trans. Audio, Speech, and Language Processing* 14: 1557–1565.

Turk, M. A. & Pentland, A. P. (1991). Face recognition using eigenfaces, *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Comput. Sco. Press, pp. 586–591.

Valente, F. (2005). *Variational Bayesian methods for audio indexing*, PhD thesis.

Valente, F. (2006). Infinite models for speaker clustering, *International Conference on Spoken Language Processing*.

Van Gael, J., Teh, Y. W. & Ghahramani, Z. (2009). The infinite factorial hidden Markov model, *Advances in Neural Information Processing Systems*, Vol. 21.

Wainwright, M. J. & Jordan, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*, Now Publishers Inc., Hanover, MA, USA.

Wooters, C. & Huijbregts, M. (2008). Multimodal technologies for perception of humans, Springer-Verlag, Berlin, Heidelberg, chapter The ICSI RT07s Speaker Diarization System, pp. 509–519.

Xiang, B., Chaudhari, U. V., Navratil, J., Ramaswamy, G. N. & Gopinath, R. A. (2002). Short-time Gaussianization for robust speaker verification, *Proceedings of ICASSP*, pp. 681–684.

Zhu, X., Barras, C., Meignier, S. & Gauvain, J. (2005). Combining Speaker Identification and BIC for Speaker Diarization, *Proceedings of Interspeech*, pp. 2441 – 2444.

Speech and Language Technologies
Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Themos Stafylakis and Vassilis Katsouros (2011). A Review of Recent Advances in Speaker Diarization with Bayesian Methods, Speech and Language Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-322-4, InTech, Available from: http://www.intechopen.com/books/speech-and-language-technologies/a-review-of-recent-advances-in-speaker-diarization-with-bayesian-methods

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen