Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach

Stephen H. Shum, *Student Member, IEEE*, Najim Dehak, *Member, IEEE*, Réda Dehak, *Member, IEEE*, and James R. Glass, *Senior Member, IEEE*

***Abstract*—In speaker diarization, standard approaches typically perform speaker clustering on some initial segmentation before refining the segment boundaries in a re-segmentation step to obtain a final diarization hypothesis. In this paper, we integrate an improved clustering method with an existing re-segmentation algorithm and, in iterative fashion, optimize both speaker cluster assignments and segmentation boundaries jointly. For clustering, we extend our previous research using factor analysis for speaker modeling. In continuing to take advantage of the effectiveness of factor analysis as a front-end for extracting speaker-specific features (i.e., i-vectors), we develop a probabilistic approach to speaker clustering by applying a Bayesian Gaussian Mixture Model (GMM) to principal component analysis (PCA)-processed i-vectors. We then utilize information at different temporal resolutions to arrive at an iterative optimization scheme that, in alternating between clustering and re-segmentation steps, demonstrates the ability to improve both speaker cluster assignments and segmentation boundaries in an unsupervised manner. Our proposed methods attain results that are comparable to those of a state-of-the-art benchmark set on the multi-speaker CallHome telephone corpus. We further compare our system with a Bayesian nonparametric approach to diarization and attempt to reconcile their differences in both methodology and performance.**

***Index Terms*—Bayesian nonparametric inference, factor analysis, HDP-HMM, i-vectors, principal component analysis, speaker clustering, speaker diarization, spectral clustering, variational Bayes.**

I. Introduction

AUDIO diarization is defined as the task of marking and categorizing the different audio sources within an unmarked audio sequence. The types and details of the audio sources are application specific, but can include particular speakers, music, background noise sources, et cetera. This paper concerns speaker diarization, or "who spoke when," the problem of annotating an unlabeled audio file where speaker changes occur (*segmentation*) and then associating the different segments of speech belonging to the same speaker (*clustering*) [1].

There exists a large amount of previous work on the diarization problem, much of which is reviewed in [1]–[3]. Because of its relative simplicity, the Bayesian Information Criterion (BIC) has served as a backbone and an inspiration for the development of a number of initial approaches involving speaker change detection and bottom-up hierarchical clustering [4], [5]. Bottom-up approaches in general, where a number of clusters or models are trained and successively merged until only one remains for each speaker, are easily the most popular in the community and consistently tend to achieve the state-of-the-art [6], [7]. A more integrated, top-down method that has achieved success is based on an evolutive Hidden Markov Model (HMM), where detected speakers help influence the detection of other speakers as well as their transitions and boundaries [8], [9]. Another approach was developed based on the "Infinite HMM," where a Hierarchical Dirichlet Process (HDP) was introduced on top of an HMM (hence, an HDP-HMM), thus allowing for up to a countably infinite number of HMM states (i.e., speakers) [10], [11]. The authors of [10] enhanced the modeling ability of the HDP-HMM by introducing a *sticky* parameter, which allows for more robust learning of smoothly varying dynamics. Subsequently, the work in [11] further extends the model to allow for explicit modeling of speaker duration.

In one sense, HDPs have become well-known in field of Bayesian nonparametric statistics, and the use of Markov Chain Monte Carlo (MCMC) sampling methods have enabled the practical application of these methods to a variety of problems [12], including diarization. However, *variational inference* is another useful technique for approximate inference that was first applied to the diarization problem in [5] and further extended in [13]. These methods, alongside the successful application of factor analysis as a front-end for extracting speaker-specific features [13], [14], serve as a basis for much of the work discussed in this paper.

Our previous work in [15] developed an approach to diarization based on the successes of factor analysis-based methods in speaker recognition [16], as well as diarization [13], [14]. Inspired by the ability of the Total Variability subspace to extract speaker-specific features on short segments of speech [16], [17], we proposed a method for performing speaker clustering directly in the low-dimensional Total Variability subspace. By evaluating the performance of our system on the same summed-channel telephone data from the 2008 NIST Speaker Recognition Evaluation (SRE), we showed that our resulting work is not

Manuscript received November 06, 2012; revised January 18, 2013 and May 02, 2013; accepted May 10, 2013. Date of publication May 22, 2013; date of current version nulldate. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steve Renals.

S. H. Shum, N. Dehak, and J. R. Glass are with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA (e-mail: sshum@csail.mit.edu; najim@csail.mit.edu; glass@csail.mit.edu).

R. Dehak is with the Laboratoire de Recherche et de Développement de l'EPITA, Paris 94276, France (e-mail: reda.dehak@lrde.epita.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2264673

only simpler than the Variational Bayes system formulated previously in [13], but can also achieve the same state-of-the-art performance.

The success achieved in [15], however, was limited to the task in which we knew there were exactly two speakers in the given conversation. To solve the diarization problem in general, we must address the setting in which the number of participating speakers is unknown *a priori*. Our work in [18] approached this problem in incremental fashion. First, we motivated the use of a spectral clustering algorithm as an alternative to the previous approach involving K-means clustering based on the cosine distance. More importantly, we adapted a heuristic from previous work applying spectral clustering to diarization and used it to determine the number of clusters (i.e., speakers) [19]. Second, we verified that there exists a symbiotic relationship between clustering and segmentation; that is, better initial segmentations yield better speaker clusters, and conversely, better speaker clusters aid in providing cleaner speaker segments. Ultimately, our system performed competitively against the state-of-the-art benchmark set by [14] on a corpus of multi-speaker telephone conversations.

This paper continues the story of [15], [18] and extends upon a number of explorations put forth in [20]. We posit that every method considered—by us and others—for speaker diarization has its advantages and disadvantages; as such, it becomes our goal to design a system that can effectively combine the advantages of different approaches and let them benefit each other with minimal supervision. To be sure, this is not a method about the combination or fusion of independently-operating systems. Rather, we extend the algorithm proposed in [18], which iteratively refines its diarization hypotheses until some form of convergence is obtained, to complement our consideration for a more probabilistic approach to speaker clustering.

There exist a number of attempts at using factor analysis-based methods for speaker diarization. The inspirations for our current saga, [13] and [14], also independently led to the work presented in [21], which uses PCA and K-means for two-speaker diarization in a way similar to our methods in [15]. Factor analysis-based features are used in [22] alongside the Cross Likelihood Ratio as a criterion for hierarchical clustering, while [23] performs clustering using PLDA as inspired by its recent success in speaker verification. Moreover, the work in [24] defines the assignment of speech segments—each represented using a factor analysis-based feature vector—to speaker clusters in terms of an Integer Linear Program. And along the lines of nonparametric methods for statistical inference, use of the mean-shift algorithm for clustering these vectors was explored in [25].

Although more detailed explanations can be found throughout the rest of this paper, we first summarize the novel contributions presented in our work below:

1) Demonstrate how applying principal component analysis (PCA) on length-normalized (i.e., cosine similarity-based) i-vectors renders them appropriate for analysis in a Euclidean space (Section IV-B-1).
2) Utilize variational inference on a Bayesian Gaussian Mixture Model (GMM) and an iterative component-death process (Section IV-B-2) to simultaneously cluster and detect the number of speakers in a given conversation.
3) Follow up on the work in [18] to further demonstrate and explain the effectiveness of iteratively optimizing segment boundaries and cluster assignments, thus taking advantage of multiple levels of information (i.e., at different temporal scales) to improve diarization hypotheses in unsupervised fashion. (Section V-B).
4) Introduce a technique to utilize the uncertainty—that is, the covariance—of an i-vector estimate, which involves drawing a number of samples from each segment's i-vector posterior distribution that is proportional to the length of the segment used to extract that i-vector (Section V-C).

In addition to presenting our proposed system in its entirety, we hope this paper can also serve to establish the notion that a factor analysis-based front-end is effective for extracting speaker-specific features from a given speech segment regardless of its length. And lastly, we hope this work can serve as an initial, though certainly not final, comparison between our proposed clustering approach using variational inference and the HDP-HMM approach using Bayesian nonparametric methods [10].

The rest of this paper is organized as follows. Section II provides some background on the Total Variability approach as a factor analysis-based front-end for extracting *i-vectors*, and Section III outlines the basic setup of our diarization system. At the theoretical heart of the paper, Section IV motivates a speaker clustering approach based on the use of PCA and a Bayesian GMM. In Section V, we outline a number of possible refinements that can be made to the system, including an extension to the iterative re-segmentation/clustering algorithm that was originally proposed in [18] and a concept known as duration-proportional sampling of the i-vector posterior distribution. The results of our experiments are analyzed and explained in Section VI, while Sections VII and VIII conclude our discussion of this work and look ahead to future possibilities.

II. Front-End Factor Analysis

At the heart of speaker diarization lies the problem of speaker modeling; logically, successful techniques in speaker modeling should also be capable of producing good results in diarization [13]. In recent years, methods in *factor analysis*, where a low-dimensional space of "factors" is used to statistically model a higher dimensional "feature space," have proven to be very effective in speaker recognition, the task of verifying whether two utterances are spoken by the same speaker [16]. We provide some intuition on how factor analysis serves as a front-end to extract relevant information from an audio sequence; more technical expositions can be found at [16], [20], [26], [27].

A. Acoustic Features

We first assume that the incoming audio has been transformed into a sequence of acoustic feature vectors. Specifically, we use raw Mel-Frequency Cepstral Coefficients (MFCCs) extracted every 10 ms over a 25 ms window. These MFCCs are 20-dimensional vectors and are the basis for our subsequent modeling. In practice, a number of variants can be used; for example,

many speaker recognition systems also include first and second derivatives into their feature vector, cepstral mean subtraction, as well as a Gaussianization feature warping step that can normalize for short-term channel effects [28]. However, in order to follow the footsteps of previous work as closely as possible, we limit our consideration to just the use of raw cepstral features, as that provided the best results in [13]. The rest of this paper assumes that all audio has been transformed into a sequence of acoustic feature vectors.

B. The Total Variability Approach

To enhance the classical method of modeling speakers using Gaussian Mixture Models (GMMs) [29], recently developed methods apply factor analysis to supervectors—a vector consisting of stacked mean vectors from a GMM—in order to better represent speaker variabilities and compensate for channel (or session) inconsistencies [16]. One such approach is Total Variability, which decomposes a speaker- and session-dependent supervector M as

$$M = m + Tw \tag{1}$$

where m is still the speaker- and session-independent supervector taken from the Universal Background Model (UBM), which is a large GMM trained to represent the speaker-independent distribution of acoustic features [29]. T is a rectangular matrix of low rank that defines the new total variability space and w is a low-dimensional random vector with a normally distributed prior $\mathcal{N}(0, I)$. The remaining variabilities not captured by T are accounted for in a diagonal covariance matrix, $\Sigma_T \in \mathbb{R}^{CF \times CF}$. The vector w can be referred to as a "total factor vector" or an *i-vector*, short for "Intermediate Vectors" for their intermediate representation between an acoustic feature vector and a supervector.

One way to interpret (1) is to see the columns of T as a limited set of directions from which M can deviate from m, the latter of which is a starting point, or bias, taken from the UBM. Ultimately, for some utterance u_S, its associated i-vector w_S can be seen as a low-dimensional summary of the speaker's distribution of acoustic features with respect to the UBM.

To avoid getting bogged down in the mathematics, we omit the details regarding the training and estimation of T and w via an Expectation-Maximization (EM) algorithm. A thorough treatment can be found in Subsection 3.3.1 of [20] as well as in [26]. For convenience throughout the rest of this paper, we use the term "i-vector extraction" to denote estimation of the posterior distribution of w (mean and covariance). Moreover, the term "i-vector" refers specifically to the posterior mean of w, while "i-vector covariance" will refer to its posterior covariance.

Lastly, the cosine similarity metric has been applied successfully in the Total Variability subspace to compare two speaker i-vectors [16]. Given any two i-vectors w_1 and w_2, the cosine similarity score is given as

$$\texttt{cos\_score}(w_1, w_2) = \frac{(w_1)^t (w_2)}{\|w_1\| \cdot \|w_2\|} \tag{2}$$

Equivalently, this means we can normalize the i-vectors by their respective magnitudes such that they all live on the unit hypersphere and the measure of the distance between two i-vectors is given by their angle.

III. System Setup

We set up the various components of our diarization system to be consistent with those of our previous work in [15], [18]. The rest of this section outlines the various parts of the system.

A. Evaluation Protocol

Before diving into the specifics, it is helpful to better understand how our system will be evaluated. Set up by NIST [30], the Diarization Error Rate (DER) is the primary performance measure for the evaluation of diarization systems and is given as the time-weighted sum of the following three error types: *Miss* (M)—classifying speech as non-speech, *False Alarm* (FA)—classifying non-speech as speech, and *Confusion* (C)—confusing one speaker's speech as from another [30]. The reference segmentation is a transcript of speech and speaker boundaries as given by the corpus. Following the conventions for evaluating diarization performance, the evaluation code ignores intervals containing overlapped speech as well as errors of less than 250 ms in the locations of segment boundaries [30]. Although overlapped speech intervals do not count in evaluating DER's, the diarization systems do have to contend with overlapped speech in performing the speaker segmentation and clustering.

B. Segmentation

In order to focus solely on the speaker confusion portion of the Diarization Error Rate (DER) and not be misled by mismatches between the reference speech/non-speech detector and our own (i.e., miss and false alarm errors), we follow the convention of previous works [13], [14] and use the provided reference boundaries to define our initial speech/non-speech boundaries. Within these boundaries, we restrict each speech segment to a maximum length of one second, and an i-vector is extracted from each. It should be noted that this rather crude initial segmentation may result in segments that contain speech from more than one speaker.

C. Clustering

The clustering stage involves grouping the previously-extracted segment i-vectors together in such a way that one cluster contains all the segments spoken by a particular speaker. And unless given *a priori*, the number of speakers (clusters) K must also be determined at this stage. Because it is known that we are strictly diarizing conversations (involving two or more participants), we require that $\hat{K} \geq 2$, where $\hat{K}$ is our estimate of K. There exist many different ways to perform clustering; Section IV provides an in-depth look at our choice of clustering method.

D. Re-Segmentation

Given a set of segments with associated cluster labels, we use the exact same re-segmentation algorithm discussed in both

[13], [15] to refine our initial segmentation boundaries. At the acoustic feature level, this stage initializes a 32-mixture GMM for each of the $K+1$ clusters (Speakers $\{S_1, \ldots, S_K\}$ and non-speech NS) defined by the previous clustering. Posterior probabilities for each cluster are then calculated given each feature vector z_t (i.e., $P(S_1|z_t), \ldots, P(S_K|z_t), P(NS|z_t)$). Pooling these across the entire conversation provides a set of weighted Baum-Welch statistics from which we can re-estimate each respective speaker's GMM. To prevent this unsupervised procedure from going out of control, the non-speech GMM is never re-trained. During the Viterbi stage, each frame is assigned to the speaker/non-speech model with the highest posterior probability. This algorithm runs until convergence but is capped at 20 Viterbi iterations, each of which involves 5 iterations of Baum-Welch re-estimation.

E. Final Pass Refinements

As in [15], we can further refine the diarization output by extracting a single i-vector for each respective speaker using the (newly-defined) segmentation assignments. The i-vector corresponding to each segment (also newly extracted) is then re-assigned to the speaker whose i-vector is closer in cosine similarity. We iterate this procedure until convergence—when the segment assignments no longer change. This can be seen as a variant of K-means clustering, where the "means" are computed according to the process of i-vector estimation detailed in [16].

IV. Speaker Clustering

Our previous work has shown that K-means clustering using the cosine distance is capable of achieving good clustering results on conversations containing any number of speakers [15], [18], [20]. Unfortunately, K-means requires as input the number of clusters to find. In [18], we adapted the use of a heuristic to estimate the number of speakers in a conversation by using a spectral clustering method, which analyzes the eigen-structure of an affinity matrix. This technique gave reasonable performance; however, its success as a heuristic only served to further inspire the development of a more principled approach.

The explorations of [20] touched upon the use of Bayesian model selection as an analog for determining the number of speakers in a conversation. Bayesian methods have the advantage of naturally preferring simpler models for explaining data. At least in theory, they are not subject to the overfitting problems which maximum likelihood methods are prone to [13].

A. The Bayesian GMM and Its Variational Approximation

Let us consider the graphical model of a Bayesian GMM as depicted in Fig. 1. Suppose each observed i-vector y_t, $t = 1, \ldots, L$, is generated by some latent speaker x_t, which is drawn according to some Dirichlet distribution (parametrized by a vector $\vec{\lambda}$) over the mixing coefficients π. By symmetry, we choose the same parameter λ_0 for each component of $\vec{\lambda}$; and as we will further discuss in Section VI-D, a small value of λ_0 will cause the resulting posterior distribution of $\vec{\lambda}$ to be influenced primarily by the observed data rather than by the prior [31].

We also introduce a Gaussian-Wishart prior to govern the mean μ_k and covariance Σ_k of the kth Gaussian component.

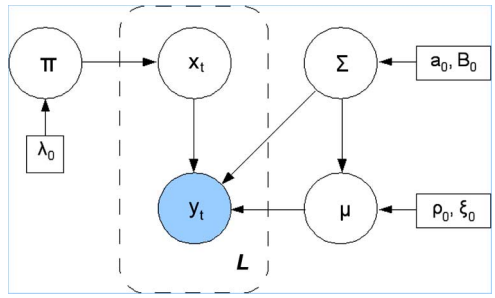


Fig. 1. A directed acyclic graphical model representing a Bayesian GMM. The dotted plate representation denotes a set of L repeated occurrences, while the shaded node y_t denotes an observation. For the parameters, $\boldsymbol{\Sigma}$ represents $\{\Sigma_1, \ldots, \Sigma_K\}$ and $\boldsymbol{\mu}$ represents $\{\mu_1, \ldots, \mu_K\}$, while the hyperparameters are shown in boxes.

Specifically, we assume $\mu_k \sim \mathcal{N}(\rho_0; \xi_0 \Sigma_k)$, thus illustrating the dependence of μ_k on Σ_k. We typically choose $\rho_0 = \mathbf{0}$; a more in-depth discussion of this model can be found in [31].

In applying this model, we ignore the time indices by which the i-vectors are created and treat each as an independent and identically distributed (i.i.d.) observation generated by some unknown (latent) speaker and attempt to identify the number of clusters (i.e., speakers) in addition to associating each i-vector (i.e., segment) with a cluster. The number of clusters can be seen as the number of mixing coefficients in π that are numerically non-trivial, though we also consider an iterative re-initialization heuristic in Section IV-B-2. And lastly, we can simply associate each i-vector to the cluster that has the highest posterior probability.

Unfortunately, the richness of Bayesian theory often renders exact probabilistic inference computationally intractable. To that end, we drew upon previous work on variational inference and applied it to the speaker clustering problem [5]. The basic idea of variational inference is to formulate the computation of a marginal or conditional probability distribution in terms of an optimization problem [12], [31]. This (generally still intractable) problem is then "relaxed," yielding a simplified optimization of a lower bound to the marginal log-likelihood[1] known as the *free energy*. To maximize this free energy, it is possible to derive an iterative Expectation-Maximization (EM) algorithm known as *Variational Bayesian EM* (VBEM). For the exact algorithmic details, we refer the interested reader to [5], [31], [32] for a more complete treatment of this topic.

B. VBEM-GMM Clustering

We turn to VBEM to perform tractable, albeit approximate, inference on a Bayesian GMM. The derivation is straightforward, and the exact parameter updates for this resulting VBEM-GMM algorithm can be found in Section 6.3 of [20] as well as in [5], [31]. Yet upon rote application of VBEM-GMM to a "bag" of i-vectors extracted from an utterance, it was clear that Gaussians are not an adequate representation for data that live on the unit hypersphere. We subsequently applied variational inference to mixtures of von Mises-Fisher distributions (Mix-vMF), but its performance did not provide sufficient gains

[1] i.e., $\log P(Y|m)$, where $Y = \{y_1, \ldots, y_L\}$ is the data and m is some given model.

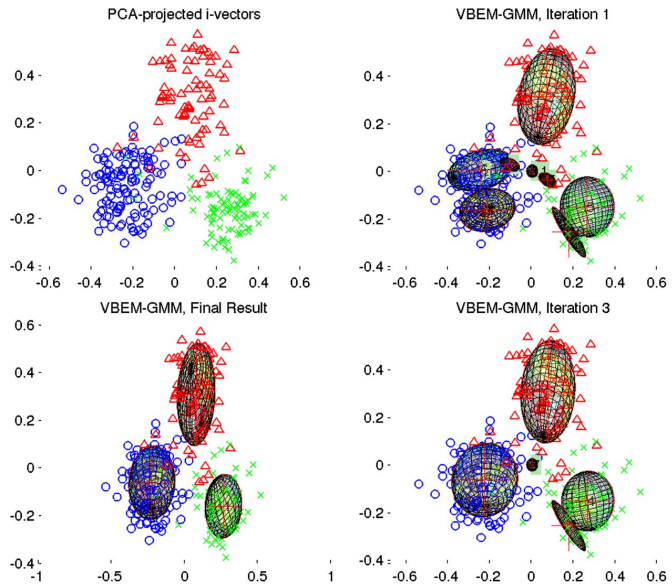


Fig. 2. (Top left) A visualization of the first two principal components of the i-vectors in a three-speaker conversation. The rest of the plots show the result of VBEM-GMM clustering after a single iteration (top right), three iterations (bottom right), and the final results (bottom left). After iterations 1 and 3, we can see that the Gaussians that do not model any significant amount of data have collapsed to the origin (i.e., their prior distribution). The clustering of the i-vectors in this utterance ultimately resulted in a DER of ∼6%.

to justify its increased complexity over the use of VBEM-GMM [20]. Ideally, there would exist some way to map our data from the unit hypersphere into a reasonable Euclidean space in which a rote application of VBEM-GMM would yield good results.

1) Dimensionality Reduction: A typical five-minute conversation is segmented into approximately 300 i-vectors, each of which lives on a 100-dimensional hypersphere. However, we should also note that each conversation in our evaluation set contains no more than seven speakers,[2] so clustering these i-vectors by speaker should not require that our data be represented in such a high dimensional space. The plot on the top left of Fig. 2 shows the first two principal components of the i-vectors in a three-speaker conversation after applying Principal Component Analysis (PCA). These points no longer lie on a unit hypersphere; rather, the Euclidean distance is now a reasonable metric for our data. Lastly, we can see that the clusters are indeed distinct despite such a limited representation, thus further supporting the validity of applying VBEM-GMM as previously mentioned.

2) An Iterative Component-Death Process: Ultimately, we would like the output of VBEM-GMM to attribute the responsibility of each i-vector to a single Gaussian; thus, how we determine the exact number of Gaussians necessary for our VBEM-GMM warrants consideration. In a so-called "birth process," we might begin with a single Gaussian and continually split components along some direction of maximal variance until the free energy is maximized [32]. Another possibility is to consider the entire range of possible cluster numbers, run VBEM on each of them, and select the result that achieves the largest free energy [31]. Empirically, we obtained our best results using a "component-death process," where we over-initialized the number of cluster components (e.g., $K_0 = 15$, although another initialization heuristic will be discussed in Section V-A) and ran VBEM. Often upon convergence of our free energy optimization, only a strict subset of those clusters ($K_1 < K_0$) will actually model any reasonable portion of the variability within the data. As such, we subsequently remove the Gaussians that are not responsible for modeling any data and randomly re-initialize VBEM with K_1 clusters. To be sure, this means we completely restart the VBEM clustering procedure as though this were the first time we have ever seen the data; the only difference is that we initialize with K_1 clusters instead of K_0. This process continues until $K_{i+1} = K_i$ for some $i \geq 0$, at which point the number of clusters has converged and we have the result of our clustering.[3]

Viewed clockwise from the top right, Fig. 2 shows the intermediate results of this clustering on the first three principal components of the same three-speaker conversation as mentioned in the preceding section. After the first iteration of VBEM-GMM (top right), seven Gaussian components remain. After the third iteration (bottom right), four components remain. At the end, we see that iterative VBEM-GMM correctly detects and clusters the three-speaker conversation accordingly (bottom left). The intermediate iterations (top and bottom right) show how the VBEM-GMM clustering free energy can get stuck in local optima, a feat not uncommon in many approximate inference methods. For this reason, the random re-initializations give the clustering method additional opportunities to find a global optimum.

V. System Refinements

The previous section explained our proposed method for speaker clustering; however, there also exist many areas in which a speaker diarization system can refine and optimize its performance. In this section, we consider a number of other possible techniques for improving our performance at the system level, the feature representation level, and the initialization level.

A. Initialization With Spectral Clustering

For our baseline experiment, the VBEM-GMM clustering method is initialized using K-means clustering (standard Euclidean distance) with $K_0 = 15$. This value of K_0 was chosen arbitrarily so as to significantly over-initialize the number of clusters without being unreasonably large. A better initialization, however, would allow the algorithm to converge faster. In [18], we obtained reasonable estimates of speaker number by adapting the use of a heuristic based on a spectral clustering algorithm [19]. The details of the algorithm itself as well as an intuitive explanation for why it works is given in [18]; here, we simply outline the steps of the algorithm needed to estimate the number of clusters.

Assume we are given n i-vectors $\{w_1, \ldots, w_n\}$ (each corresponding to a speech segment ≈1 sec in length).

[2]To be sure, this fact is not used as an input to our diarization system.

[3]We should admit that this is not at all a fully Bayesian solution, nor did we intend for it to be. We chose to use a Bayesian GMM and, hence, efficient variational inference, because in contrast to maximum likelihood, such methods are less likely to overfit the data when presented with an over-initialization of the number of clusters.

Form the affinity matrix $A \in \mathbb{R}^{n \times n}$, where $A_{ij} = \exp(-d(w_i, w_j)^2/\sigma^2)$ when $i \neq j$ and $A_{ii} = 0$. Here, $d(w_i, w_j) = 1 - \texttt{cos\_score}(w_i, w_j)$, where cos_score is given by (2). For reasons explained in Section 4.1 of [18], the scaling factor σ^2 is set to be 0.5. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row, and construct the matrix $L = D^{-(1/2)}AD^{-(1/2)}$ [33].

It was seen experimentally that the n sorted eigenvalues of L, say $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, exhibit exponential decay and that the number of speakers in a conversation correspond consistently to when the gradient of these eigenvalues exceeds some threshold θ. As such, to determine the number of clusters, we can fit, in a way that minimizes the mean squared error, a smooth exponential $\exp(-\alpha k)$ to Λ, where $k = 1, \ldots, n$ and $\alpha \in [0.1, 10]$. We then take $\hat{K}$ to be the smallest value whose derivative $-\alpha \exp(-\alpha \hat{K}) \geq \theta$ [18].

Because of the non-increasing nature of our iterative VBEM-GMM clustering (i.e., $K_{i+1} \leq K_i$), we need to ensure that the spectral clustering-based initialization is, with high probability, greater than the actual number of speakers in the conversation K^*. Indeed, we want a more informed initialization than $K_0 = 15$, but it would be far worse to initialize the clustering with an underestimate that forces some clusters to model speech from more than one speaker, thereby irreversibly corrupting our speaker models. Upon looking at the error distribution of $\hat{K}$—the number of speakers estimated via spectral clustering—over development data, we introduce a bias such that our initialization $K_0 = \hat{K} + \lceil 3\sigma_{\hat{K}} \rceil$, where $\sigma_{\hat{K}}$ is the standard deviation of the error distribution of $\hat{K}$ and the ceiling function $\lceil \cdot \rceil$ ensures that K_0 is an integer.

B. Iterative Optimization

It was shown in [18] that when the number of speakers needs to be estimated, improved results are obtained via an iterative optimization procedure, which alternates between clustering and re-segmentation until the diarization hypothesis converges. Similar to the notion of giving the iterative VBEM-GMM clustering method more opportunities to find a global optimum, the iterative optimization procedure gives the system more opportunities to re-estimate the number of speakers using (hopefully) cleaner and more refined speech segments. We follow the explanation provided in [18] to reiterate the intuition behind this idea.

The use of factor analysis for speaker diarization allows us to take advantage of multiple levels of speaker information. I-vectors are designed to provide information specific to speaker (and channel) identity, which is important for clustering; however, the effectiveness of an i-vector is proportional to the length of the speech segment from which it is extracted, thus it is not as well-suited for issues requiring finer temporal resolution (e.g., speaker change detection). By contrast, lower-level acoustic features such as MFCCs are not quite as good for discerning speaker identities, but can provide sufficient temporal resolution to witness local speaker changes and segment boundaries.

To that end, we formulate an algorithm that optimizes both segmentation boundaries and segment cluster assignments in iterative fashion. More specifically, we can alternate between VBEM-GMM clustering (done at the i-vector level) as described in Section IV-B and applying the re-segmentation method (done at the acoustic feature level) as described in Section III-D until successive diarization hypotheses "converge." In general, this iterative concept was proposed initially in [34] and then adopted by other systems in practice [6], [9]. Our exact approach was inspired by the work in [13]—they began with a crude initial segmentation and ran factor analysis-based clustering followed by Viterbi re-segmentation and then a second pass of the clustering (using the new segmentation) to obtain a final diarization hypothesis—we have simply formalized this idea and introduced the notion of convergence. Let us approximate a "distance" between two diarization hypotheses D_1 and D_2 by running it through a diarization evaluation script as provided by NIST [30]. Then we can define a "convergence" to be when this error rate (i.e., $\mathrm{DER}(D_1, D_2)$) between the hypotheses from the previous two iterations is below some threshold γ. In our experiments, we set $\gamma = 8\%$ and allow a maximum of 20 total iterations.[4] These values were set to optimize a combination of both system performance and run-time on the development set. Ultimately, our test results required an average of 3.7 iterations per conversation; the numbers varied widely by conversation,[5] however, and were independent of both the number of speakers present and the resulting DER.

It should be noted that the re-segmentation output from Section III-D includes both segment boundaries and corresponding cluster assignments. During this iterative optimization process, however, the assignment labels from the re-segmentation output are not provided as input to the clustering stage—only the segment boundaries are considered. Lastly, the number of speakers is also re-estimated at the start of each clustering stage. By requiring that the clustering and re-segmentation steps are run in this completely disjoint fashion, we ensure that information from different temporal resolutions is used only for its designed purpose; that is, only information at the i-vector level will be used for the clustering of segments, and only information at the acoustic feature level will be used to determine segment boundaries.

C. Duration-Proportional Sampling

Our discussion thus far has been restricted to the use of i-vectors as point estimates. During clustering, all i-vectors are treated as independent, identically distributed i.i.d. samples from some underlying distribution. This assumption, however, is not necessarily true. For example, a segment that is five seconds long gives a much better representation of the speaker than a segment 0.5 s in length; yet, both segments yield i-vectors of the same dimensionality and are treated equally during clustering.

Recall from Section II-B that the "i-vector" is merely the posterior mean of w as given by (1). There is also an associated posterior covariance of w, which we termed "i-vector covariance," whose determinant (i.e., "volume") is actually inversely

[4]Setting tighter convergence threshold, i.e., smaller values of γ, brought little improvement to overall performance at the expense of significantly increased computation time.

[5]Standard deviation $= 4.5$ iterations, and 4% of diarization hypotheses did not converge after 20 iterations.

proportional to the number of acoustic frames used to estimate the posterior distribution [20], [26]. Thus, the longer the segment used to obtain an i-vector, the smaller its posterior covariance (uncertainty), and the more robust the speaker estimate.

To make use of durational and covariance information, we consider the following sampling scheme. For a given i-vector w and its covariance $\mathrm{cov}(w)$, we draw a number of samples n from this distribution proportional to the time duration t of the segment used to estimate $\mathcal{N}(w, \mathrm{cov}(w))$. This technique makes use of durational information in two ways: (a) a shorter segment results in relatively fewer i-vector samples, and (b) a shorter segment results in a covariance $\mathrm{cov}(w)$ that is relatively large, thus its samples will range more widely. Conversely, a long segment will have a lot of samples concentrated in a small part of the space. This takes advantage of the difference in uncertainty between segments of different length by increasing the relative importance of longer, more reliable segments for the estimation of our respective speaker clusters.

In our experiments, we sample from our i-vectors at a rate of four samples per second of conversation; our original approach using one-second segments resulted in approximately one i-vector per second. Similar to the convergence criterion γ in Section V-B, this sample rate was chosen to optimize between increased system performance and run-time, as a higher sampling rate requires more computation for the clustering algorithm. Given these samples, we apply PCA and put them through the VBEM-GMM clustering as usual, resulting in an assignment of each sample to some corresponding GMM cluster. We then assign a cluster to the respective i-vector from which each of these samples was drawn by picking the GMM cluster that represents the majority of its samples.

VI. Experiments

In order to use the same telephone-based Total Variability framework from [15], [18] and utilize the state-of-the-art results from [14] as a benchmark for comparison, we evaluate our system on the 2000 NIST SRE subset of the multilingual CallHome data, a corpus of multi-speaker telephone conversations. This amounts to 500 recordings, each 2–5 minutes in length, containing between two and seven participants [35]. Also associated with this test set is a development set,[6] which consists of 42 conversations, each at least five minutes in length, featuring between two and four speakers. With the exception of Japanese, all the languages present in the CallHome test set are also represented in the development set. Table I provides a summary of the CallHome corpus, including both the development set (in parentheses) and the test set, broken down by number of speakers and language spoken. We will break down our results to show diarization performance on conversations involving the different numbers of speakers.

TABLE I
Summary of CallHome Corpus Broken Down by Number of Participating Speakers and Language Spoken. The Numbers in Parentheses Represent Those in the Development Set, While the Values Not Enclosed in Parentheses Represent Those in the Test Set

	NUMBER OF SPEAKERS						
	2	**3**	**4**	**5**	**6**	**7**	**TOTAL**
Arabic	50 (4)	28 (4)	10 (2)	3	4		**95** (10)
English	49 (4)	7 (2)					**56** (6)
German	52 (4)	12 (2)	3				**67** (6)
Japanese	53	10	2	3			**68** (-)
Mandarin	47 (4)	50 (4)	18 (2)	2	1		**118** (10)
Spanish	52 (5)	29 (3)	10 (2)	2	1	2	**96** (10)
TOTAL	**303** (21)	**136** (15)	**43** (6)	**10**	**6**	**2**	**500 (42)**

A. Implementation Details

We obtain our i-vectors using the same Total Variability matrix T of rank 100 that achieved the best reported results in both [15] and [18]. This matrix was trained from a gender-independent UBM of 1024 Gaussians built on 20-dimensional MFCC feature vectors without derivatives. Both the UBM and T were built using the Switchboard (English) and Mixer (multilingual) Corpora; the latter was used during the 2004, 2005, and 2006 NIST SREs. Overall, these data include over 1000 hours of speech from a variety of different languages and, for the most part, match the data used to train the models in [13].

A primary goal of designing this system was to require the tuning of as few parameters as possible. Of course, some were unavoidable—for example, defining the threshold for diarization hypothesis convergence (Section V-B), or estimating the bias term in the spectral clustering initialization of the number of speakers (Section V-A)—but even those required only coarse adjustments. The Bayesian structure of our speaker clustering method further limited the number of hyperparameters that require consideration; in fact, the only exception was choosing the Dirichlet concentration parameter on the distribution of mixture weights for VBEM-GMM.

There exist a number of methods for choosing hyperparameter values. To obtain an empirical prior, Section 3.1.3 of [5] outlines an EM-like algorithm that converges on values of the hyperparameters which maximize the variational free energy. An even more principled way to approach this would be to assume a prior distribution on the hyperparameters and sample them accordingly [10]. For simplicity, we chose to use the hyperparameters that achieve the best result (in the DER sense) on the associated development set. We should note immediately, however, that there is a significant mismatch between the development set and the test dataset; in particular, test conversations feature up to seven speakers and can be as short as two minutes. We demonstrate in Section VI-D that our proposed methods are relatively robust to this mismatch; the subsequent results we report in Sections VI-B and VI-C are based on the parameters that achieve the best DER performance on the development dataset.

We make use of an existing MATLAB implementation of VBEM-GMM provided in [36] and build our VBEM-GMM clustering as described in Section IV-B. We run PCA on a per-utterance basis using our length-normalized i-vectors and keep only the first three principal components to perform clustering in the manner depicted by Fig. 2. There exist many ways to refine this method of dimensionality reduction; however, that is beyond the scope of this paper, and we postpone further discussion of this topic until Section VIII.

[6]We would like to thank Craig Greenberg of NIST for making this available.

B. System Comparisons

The plot at the top of Fig. 3 shows the results of our VBEM-GMM clustering in comparison with our proposed system refinements as well as the state-of-the-art benchmark set on this task in 2008 by Castaldo, *et al.* [14], which we show in black. Shown in magenta are the results of our initial baseline system, in which we implement the VBEM-GMM clustering ($K_0 = 15$) on 3-dimensional, PCA-projected, and length-normalized i-vectors. After clustering, we run a single iteration of the re-segmentation algorithm discussed in Section III-D and finish with a set of final pass refinements (Section III-E). We can see from the plot that our baseline achieves results similar to that of [14] on conversations involving four or more speakers. However, our system does not perform as well on conversations containing only two or three speakers, which make up the overwhelming majority of the dataset. A similar story unfolds when we initialize using the spectral clustering heuristic discussed in both Section V-A and [18]. Shown in blue, this method of initialization provides slightly better results in the two-speaker case and similar results otherwise compared to the initial baseline system ($K_0 = 15$).

1) Regarding Diarization Error: One of the reasons that can be attributed to this large error deviation is that of over-estimating the number of speakers. This effect is most prominent in the case of two-speaker conversations. For example, suppose a two-speaker conversation is segmented such that all the segments attributed to speaker A are assigned to cluster I, but the segments attributed to speaker B are assigned arbitrarily to clusters II and III. On one hand, our diarization system has done a reasonable job of distinguishing between two speakers; on the other, it has failed to realize that two separate clusters (II and III) actually belong to one speaker. Such an error is forgivable and, in fact, can be easily remedied in a post-processing step by the use of a more powerful speaker recognition system, such as in [16]; conversely, it would have been much worse to combine two different speakers into a single cluster. Unfortunately, the less-forgiving Diarization Error Rate (DER) penalizes both types of errors equally heavily: If cluster I represents half of the conversation time and each of clusters II and III represent a quarter of the conversation time, then the DER would be 25%, which is a bit unreasonable given that each of these clusters are nevertheless pure representations of exactly one speaker.

In light of this, it might be reasonable in subsequent work to consider another performance metric for judging our methods, such as Average Cluster Purity (ACP) [5]. This, of course, has yet its own set of advantages and disadvantages—namely that we can obtain perfect cluster purity (i.e., 100%) by letting each segment be its own cluster—but for the sake of providing additional perspective in contrast to DER, we display the ACP of our diarization results at the bottom of Fig. 3. In general, if a particular cluster represents the speech of N different speakers speaking t seconds of speech, then its purity is the proportion of t corresponding to the speaker that speaks the most in that cluster. Whereas a one-to-one mapping is required in the computation of DER, cluster purity allows for many clusters to represent a single speaker. We compute ACP by taking a time-weighted average of each cluster's purity such that a cluster representing a larger proportion of the conversation will contribute more to the ACP.

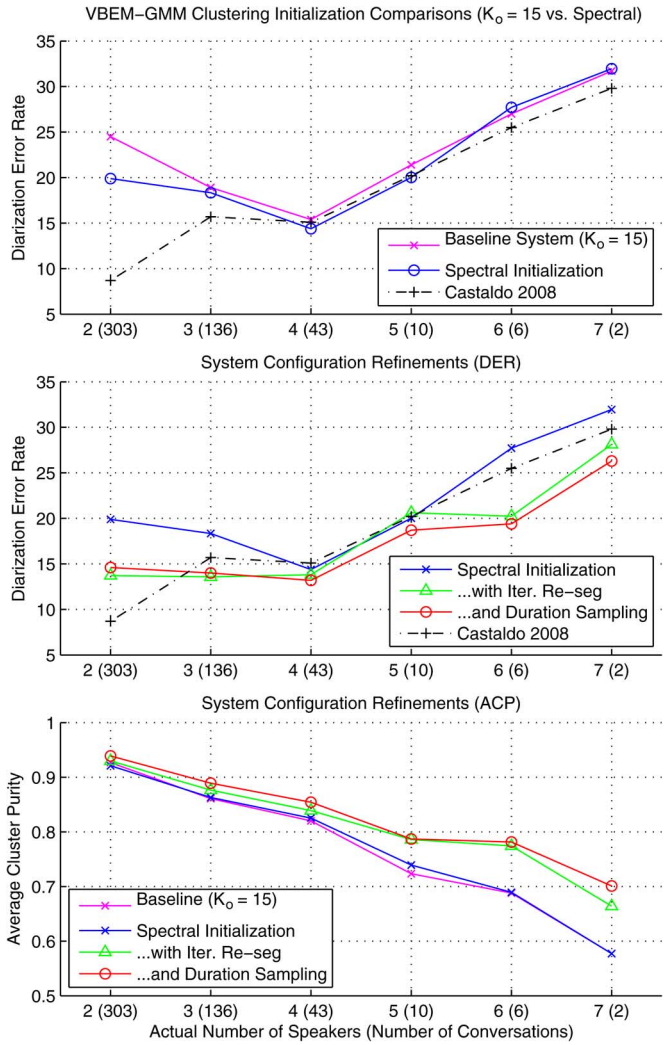


Fig. 3. (Top) results comparing the baseline initialization of VBEM-GMM using $K_0 = 15$, in magenta, with an initialization using the spectral clustering heuristic described in Section V-A, in blue. (Middle) results obtained after incorporating the various system refinements proposed in Section V. In blue is our baseline that initializes VBEM-GMM using the spectral clustering heuristic (same as the plot on top). We show the state-of-the-art benchmark results from [14] in black. (Bottom) for each of the systems whose DER results we show above, we also show its Average Cluster Purity (ACP) using the same line color coordination and similar marker type.

2) Evaluating System Refinements: Confirming our hypothesis from Section V-A, the spectral clustering initialization gives slightly better results than the baseline initialization with $K_0 = 15$ speakers. Its most prominent effect was on two-speaker conversations, where a more informed initialization gives the VBEM-GMM clustering a better chance of properly detecting two speakers, thus driving down the DER. Our subsequent experiments use the spectral initialization as the new starting point (baseline).

The two lower plots on Fig. 3 show the results obtained after incorporating the various system refinements proposed in Section V. We can see that the iterative re-segmentation/clustering optimization (Section V-B) has a mostly positive effect

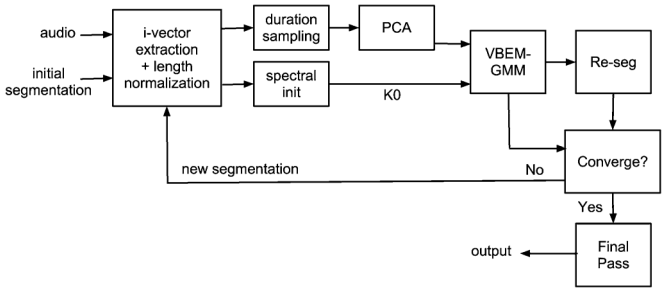


Fig. 4. Final system diagram.

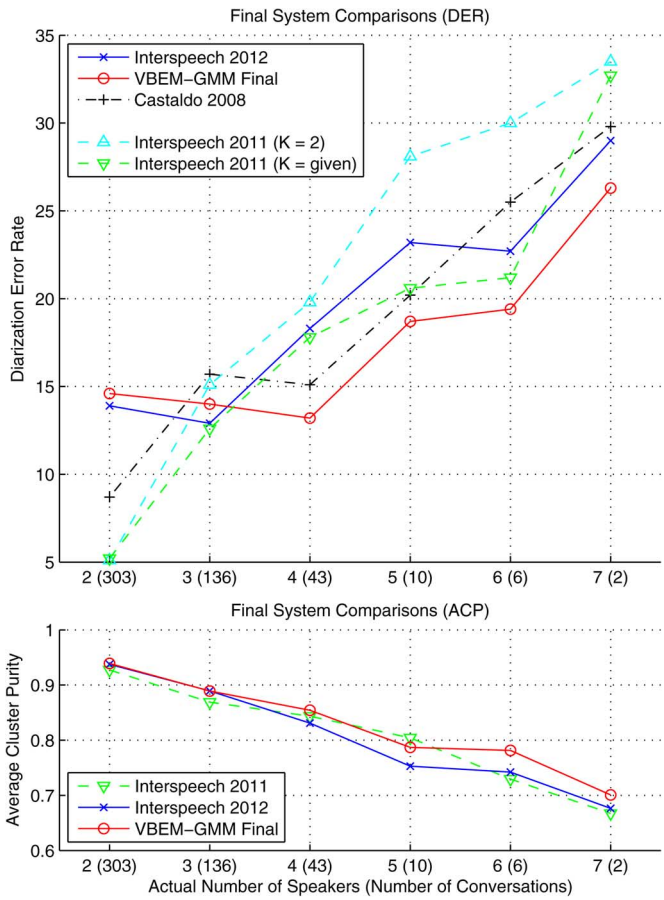


Fig. 5. (Top) final results comparison between our current final system (in red), the system from our previous work in [18] (in blue), and the state-of-the-art benchmark system proposed in [14] (in black). Also shown are results from the initial system in [15] that always assumes the presence of only two speakers ($K = 2$) from our initial work (in light blue) as well as results from the same system where we provide the number of speakers, i.e., K is given (in green). (Bottom) we also provide a comparison between the resulting average cluster purity of these systems. Note that the points labeled "Interspeech 2011" correspond to the case in which the number of speakers K is given.

on both DER and ACP, as does the duration-proportional sampling (Section V-C), which we implemented at a rate of four (i-vector) samples per second. Incorporating all of these system refinements gives our best overall performance.

C. Final System

To facilitate understanding, a block diagram of our final system is shown in Fig. 4. Given some initial speech/non-speech segmentation, this system extracts length-normalized i-vectors and then, in parallel with estimating the number of clusters using our modified spectral clustering heuristic, performs both duration-proportional sampling and a PCA-projection to three dimensions. At this point, we run VBEM-GMM clustering and Viterbi re-segmentation. This process iterates until successive diarization hypotheses meet our convergence criterion, after which we run the hypothesis through a final pass of refinements as discussed in Section III-E to obtain our final result.

We also compare this final system to the system proposed in [18], where the setup is analogous to the one proposed in this paper; the difference is that our previous method used only the spectral clustering heuristic to determine the number of speakers and K-means (based on the cosine distance) to obtain the actual cluster assignments without the need for dimensionality reduction via PCA. Otherwise, the iterative re-segmentation/clustering optimization and final pass refinements are common to both approaches. Fig. 5 shows this comparison, and we can see that our current system in question (shown in red) provides a noticeable improvement from our previous approach in [18] (in blue) on conversations involving four or more speakers while displaying no substantial difference in performance on conversations involving two or three speakers. Similarly, our current system performs better than the state-of-the-art benchmark (in black) in all settings except for conversations involving just two speakers.[7]

Table II summarizes the hyperparameters that were used to generate our final results, while the bottom of Fig. 5 shows the resulting ACP of these final systems. The green-dashed line with inverted triangular markers labeled "Interspeech 2011" corresponds to the case in which the number of speakers K is given, and we can see that our final system (red line, circular markers) also provides the purest clusters overall. Across all the conversations in the CallHome test set, our Interspeech 2011 system provides an ACP of 89.8%, the Interspeech 2012 system provides an ACP of 90.8%, and the system we propose in this paper gives an overall ACP of 91.2%.

TABLE II
SYSTEM PARAMETERS AND THEIR VALUES USED TO OBTAIN THE RESULTS SHOWN THROUGHOUT THIS PAPER. UNLESS EXPLICITLY STATED, NONE OF THESE VALUES WERE OPTIMIZED ON THE CALLHOME TEST SET; THEY WERE EITHER INHERITED FROM PREVIOUS WORKS [13], [15], [18] OR OPTIMIZED ON THE CALLHOME DEVELOPMENT SET

Parameter Name	Value
Acoustic features	20-dim raw MFCCs
Universal Background Model	1024 Gaussians
i-vector dimensionality	100 dimensions
PCA dimensionality	3 dimensions
Bayesian GMM Dirichlet concentration parameter, λ_0	1e-18 (dev-best); 1e-6 (test-best)
Threshold for spectral clustering eigenvalue decay, θ	0.001
Standard deviation of spectral clustering speaker estimation error, $\sigma_{\tilde{K}}$	0.64
DER convergence threshold, γ	8%
Maximum iterations	20
i-vector sampling rate	4 samples/second

[7]We discuss the results obtained using "Interspeech 2011 ($K = 2$)" (dashed light blue line, upright triangular markers) as well as "Interspeech 2011 (K=given)" (dashed green line, inverted triangular markers) in Section VI-E.

With regard to computational requirements, we did not run a controlled benchmark test on the amount of time it took to complete an evaluation, nor did we take any measure to optimize the performance of our implementation to ensure its efficiency. As such, our mix of MATLAB, Perl, and Bash scripts required around 60 hours to evaluate the CallHome test set (500 recordings, 2–5 minutes each ≈30 hours) on a single quad-core machine. We should note, however, that while the methods discussed here are designed more for optimal performance than to obtain a lightweight diarization system, they can be tuned and modified to work much more quickly. For example, the main computational bottleneck lies in the convergence of our iterative optimization scheme, which can—and in 4% of conversations, did—require up to 20× as long as a one-pass, sequential diarization system. For each subsequent iteration, new i-vectors need to be extracted for each segment of speech; a pairwise affinity matrix and its eigenvalues need to be computed for spectral analysis; then VBEM-GMM clustering is run followed by the entire Viterbi re-segmentation process on acoustic features. In general, each of these individual steps can run reasonably quickly, but the fact that iterative optimization may require these steps to repeat a variable number of times inevitably increases the computation time significantly. We reserve for future analysis the effect of relaxing our DER convergence threshold, γ (Section V-B), on the resulting system performance.

D. Parameter Robustness

To evaluate the robustness of our system, we explore the sensitivity of our test results to different values of the Dirichlet concentration parameter, λ_0. This parameter quantifies a prior belief of how evenly the responsibility should be shared amongst the various components of our mixture model. In particular, letting $\lambda_0 \to 0$ will yield clustering solutions in which more and more of the mixing coefficients are zero; that is, more and more mixture components will not model any data. This makes sense for our purposes, as we deliberately over-initialize the number of clusters in order to prune them away via an iterative component-death process as described in Section IV-B-2.

To arrive at the results shown in Figs. 3 and 5, we picked the value of λ_0 that achieved best results (in the DER sense) on the CallHome development set. Upon experimenting with different values of λ_0, however, we observed that the resulting differences in performance on the CallHome test set were mostly minor and insignificant. Fig. 6 shows these results, which suggest that our proposed system is not terribly prone to overfitting on development data and can potentially generalize well to other test sets, though further experimentation would be required before we can formalize this claim.

E. Discussion

Admittedly, it is rather frustrating that we are unable to do better on two-speaker conversations. Incidentally, our final system is based off of the same setup that obtained state-of-the-art results in the task of two-speaker diarization on 2008 NIST SRE data, where the number of speakers is given. In [15], our system performed at least as well as each of the systems described in [13], one of which was actually the same system (Castaldo 2008 [14]) that we use as our benchmark in this paper.

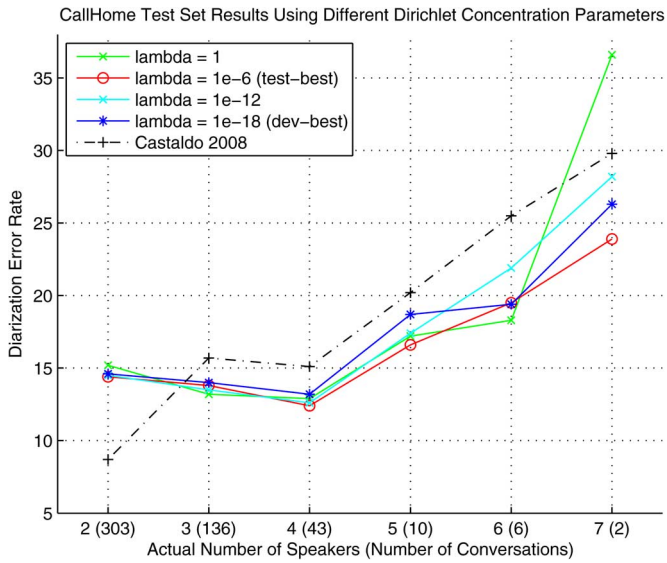


Fig. 6. Results on CallHome test set using different values for the Dirichlet concentration parameter, λ_0.

There is no easy way to reconcile the inability of our final system to match the performance of our two-speaker diarization system in [15]. One possible explanation is that even despite sweeping across different Dirichlet concentration parameter values on the test set, we seem prone to over-estimating the number of speakers when diarizing two-speaker conversations. We discussed previously in Section VI-B-1 the harshness of the DER metric on over-estimating the number of speakers. Further analysis shows that on the 136 out of 303 two-speaker conversations in which we correctly detected two speakers, our DER is in fact lower than the DER reported in [14] (6.5% vs. 8.7%).[8] And when the system is given that there are exactly two speakers in the conversation, the DER drops even further to 4.3% [18]. These are, of course, unfair comparisons; however, they do provide some measure of consolation for the seeming inconsistencies that we see in our final results.

This brings to bear the question of what results we would obtain if our system were to simply assume that every conversation contained exactly two speakers. Judging by the distribution of speakers per conversation on the last row of Table I, it is clear that two-speaker conversations make up the majority of this test set. Just to get a better sense of the baseline standard for our proposed techniques, we run the exact system from [15], which obtained state-of-the-art diarization results on two-speaker conversations, on the CallHome data and show their results (in light blue) in Fig. 5. To be sure, this system extracts an i-vector for each speech segment, runs K-means clustering using the cosine distance, and then undergoes a single pass of the Viterbi re-segmentation algorithm (without iterative optimization) before going through a set of final pass refinements. Not surprisingly, this method achieves results on two-speaker conversations (5.1% DER) that approach the 4.3% DER attained

[8]Within this subset of two-speaker conversations, we detected three speakers 97 times, four speaker 52 times, five speakers 12 times, and six or more speakers 6 times.

after incorporating the iterative optimization of segmentation and clustering [18]. What is more interesting, however, is that this system also attains reasonable results on conversations involving more than two speakers. This could be evidence that telephone conversations are often dominated by only two participants. For good measure, we also show the results (in green) obtained by the same system from [15], but in which the provided value of K corresponds to the actual number of speakers in the conversation (i.e. an "oracle" experiment). We can see that our proposed methods—both from [18] and those in this paper—do significantly better than the two-speaker assumption, and in particular, the techniques described in this paper also outperform the "oracle" system.

VII. Further Analysis

The diarization task in which we are given the number of participating speakers is wholly different from the task in which the number of participating speakers needs to be estimated. Our state-of-the-art performance on two-speaker diarization in [15] really only served to further validate that factor analysis, and i-vectors in particular, is a viable front-end for extracting utterance-specific features from the short speech segments featured in diarization. From there, it is in the way that these features are processed that truly defines the effectiveness of the diarization system.

A. Towards Temporal Modeling

We pointed out in [18] that the benchmark system in [14] is, whether intentional or incidental, actually designed to take advantage of the structure of telephone conversations. In particular, most speaker turns over the telephone involve no more than two participants at any given time. The system in [14] processes these calls in causal fashion, working on 60-second slices and assuming that each slice contains no more than three speakers. Given the nature of the data, this makes sense; except for the rare use of speakerphones, only during these relatively infrequent "hand-off" scenarios would a third speaker even exist in any particular slice of the conversation.

By contrast, our algorithm sees and processes an entire utterance at once and performs clustering without any regard to the potentially restrictive temporal dynamics of a telephone conversation. This so-called "bag of i-vectors" approach may be slightly more general in its ability to handle scenarios in which four or more speakers appear in any 60-second slice of the conversation (a hypothesis not tested for in [14]); however, it also has the inherent disadvantage that it is more prone to missing speakers that, say, only participate in a very short snippet of the conversation. This refers back to the problem of data sparsity, or inadequate cluster representation, as mentioned in Section V-C. One way to overcome this might be to modify our approach to process the data incrementally, where clustering is run on shorter, say 60-second, slices of conversation before linking clustered slices across an entire utterance. For speakers that only participate in a limited portion of the conversation, the shorter slice-based processing gives them the opportunity to be better-represented when we cluster the slice in which they are (relatively) more active. Yet another, possibly more principled, way to approach this issue might be to model temporal dynamics—including the entrance and exit of a particular speaker—directly from the conversation.

B. A Sticky HDP-HMM

The sticky HDP-HMM is a Bayesian nonparametric method for statistical inference that achieved state-of-the-art results in meetings diarization on the NIST Rich Transcription (RT) 2004–2007 database [10]. The authors leverage the "importance of temporal dynamics captured by the HMM" as a way to improve their baseline results obtained from a "Dirichlet Process mixture-of-Gaussians model (ignoring the time indices associated with the observations)," which is analogous to our Bayesian GMM [10]. Because our work utilizes improved speaker modeling using a factor analysis-based front-end (instead of smoothed acoustic features; i.e., MFCCs averaged over 250 ms [10]), we were interested to see what further gains could be obtained by incorporating temporal modeling with i-vectors. Moreover, one of the fundamental limitations of an HMM in general is that observations are assumed conditionally i.i.d. given the state sequence [10]. Even though i-vectors still violate this property somewhat, we believe that they are better suited than acoustic features (i.e. less temporally correlated) to the conditional independencies assumed by the HMM generative model. The details of the HDP-HMM model itself as well as a method to perform efficient blocked Gibbs sampling are thoroughly explained in [10].

Using the implementation provided by [37], we explore the performance of the sticky HDP-HMM on i-vectors extracted from the CallHome evaluation set by replacing the VBEM-GMM module from our system depicted in Fig. 4 with the sticky HDP-HMM. For proper and comprehensive comparison with our current and previous results, we optimized the associated hyperparameters over both the development set and the test set in the same manner as described in Section VI-A. Fig. 7 shows the results in terms of both DER and ACP.

The sticky HDP-HMM seems to provide a significant improvement over both our VBEM-GMM and Castaldo's [14] systems on two-speaker conversations. Such an outcome, however, is also attributed to the fact that we enforce a minimum of two detected speakers, as mentioned in Section III-C. If the sticky HDP-HMM clustering (or, similarly, VBEM-GMM clustering) returns just one speaker, the system backs off to K-means clustering where $K = 2$. Out of 303 two-speaker conversations, the initial sticky HDP-HMM clustering returned one speaker for 106 of them and returned two speakers for 143 conversations. That said, because this back-off technique is common to both the HDP-HMM and VBEM-GMM approaches, it seems that—in spite of choosing hyperparameters for optimal DER—the VBEM-GMM approach is prone to overestimate the number of speakers, while the HDP-HMM approach tends to underestimate.

As for conversations involving other numbers of speakers, the sticky HDP-HMM is competitive, in the DER sense, with the VBEM-GMM on conversations involving exactly three speakers, but results start to deteriorate for both DER and ACP as the number of speakers increases. Lastly, there seems to be a discrepancy in test performance between the different hyperparameters that optimize the development set and those that

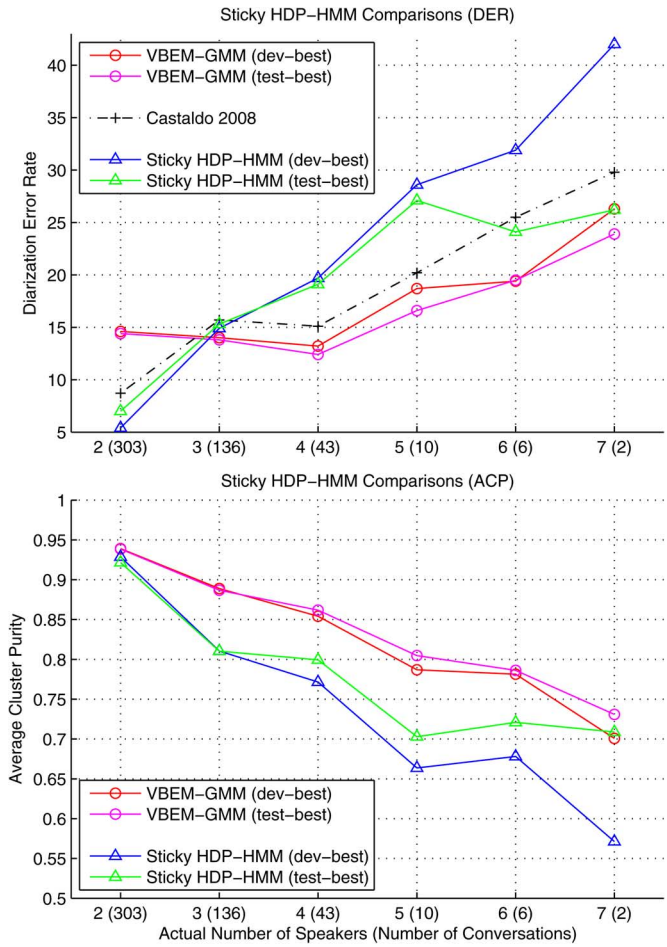


Fig. 7. (Top) a comparison between our dev- and test-best VBEM-GMM systems (in red and magenta, respectively), the state-of-the-art benchmark system proposed in [14] (in black), and the dev- and test-best sticky HDP-HMM systems proposed in [10] (in blue and green, respectively). (Bottom) corresponding results in terms of average cluster purity.

optimize the test set. Because this difference involves only a small subset of the evaluation, however, it should be considered minor. Nevertheless, this once again highlights a fundamental mismatch between the development set and the test set, and perhaps the hyperparameters of the sticky HDP-HMM are more sensitive to the mismatch than our VBEM-GMM parameters.

Further exploration on the topic of Dirichlet processes suggests that the dependent Dirichlet process (DDP) might be an appropriate way to model the temporal constraints of the CallHome telephone data [38]. In this method, a DDP changes according to a Markov chain, where the Dirichlet process drawn at any particular time interval is dependent on the Dirichlet process of the previous interval. In this way, each Dirichlet process models only a local or limited portion (i.e., slice) of the conversation. From one slice to the next, the participation of speakers can be introduced, removed, or modified (e.g., the transition from a monologue to an open discussion). Thus, if a telephone conversation were processed in slice-based fashion [14] as described in Section VII-A, the DDP provides an elegant framework that allows for the modeling of handset "hand-offs" and conversation dynamics. We believe this could be a potential avenue for future work.

VIII. Conclusion

In this paper, we have continued the story of our previous work in developing a system for speaker diarization based on a factor analysis-based front end [15], [18], [20]. Our final system contains traces of inspiration from pioneering works in diarization using factor analysis [14], variational Bayesian inference [5], and both in combination [13]. We have obtained results that are comparable to the current state of the art, and more importantly, we have demonstrated such performance with the use of well-known, off-the-shelf machine learning techniques. From the i-vector and its cosine similarity metric to PCA and VBEM-GMM clustering to the use of a spectral initialization and an iterative optimization process, each of our methods were chosen not only to exploit various properties of the data, but also to complement each other in the spirit of the diarization task itself. What results is a system that is mostly unsupervised and reasonably robust.

We also compared our approach to a Bayesian nonparametric method that incorporates temporal modeling in the form of a sticky HDP-HMM [10]. This was an initial and exploratory attempt at replacing smoothed acoustic features with i-vectors and modeling the temporal dynamics explicitly. Despite the tendency to underestimate the number of participating speakers, this approach achieved very competitive performance on conversations involving small numbers of speakers. Nevertheless, this warrants a more in-depth analysis to better compare these methods.

We realized, for all systems, that the diarization hypothesis that attains the best DER is not always the one that correctly detects the number of speakers. That is, forcing the system to detect an exact number of clusters would often have a detrimental effect on the DER (except, apparently, in the case where there are only two speakers!). One reason for this goes back to the problem of inadequate cluster representation, where a speaker's contribution might be so limited that enforcing an exact number of clusters ends up splitting another speaker into two clusters. Because the relative amount of participation amongst present speakers in each test conversation lacked the sort of uniformity or predictability that would have made for an appropriate evaluation of accuracy in detecting the number of speakers, we instead focused the efforts of this paper towards optimizing our system for minimal DER.

There are still many ways in which we can improve and refine the steps to our approach. For one, we do all our VBEM-GMM clustering using just the first three principal components of our i-vectors. This initial choice of dimensionality was primarily for purposes of visualization; however, using a different number of dimensions did not change results significantly. Further investigation on dimensionality choice as well as other potential methods for dimensionality reduction should yield a more insightful understanding and, hopefully, more fruitful results [39]–[41]. Second, our hyperparameters were determined by trying a number of different values and observing the resulting performance on some development set. It would be nice to see the result of following a more principled and "Bayesian" approach to setting our prior hyperparameters as mentioned in Section VI-A and more thoroughly discussed in [5], [10], [31].

Finally, our evaluation was restricted to the diarization of telephone conversations. Much of the current work in diarization has moved into the realm of broadcast news and meetings, such as those of the NIST RT database [4], [10]. The reason we limited ourselves to telephone data was to fully exploit the effectiveness of our data-driven factor analysis-based front-end, which requires ample background data to build. But as our ability to model microphone data approaches the current standard of telephone data modeling [42], we look forward to extending our methods to the diarization of meetings and seeing whether the proposed approaches discussed in this paper can achieve equally good performance.

ACKNOWLEDGMENT

We would like to thank the editor and reviewers for their helpful comments and feedback in the development and refinement of this work. We would also like to thank Douglas Reynolds for his insightful discussion and helpful advice through the years.

REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[2] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Commun.*, vol. 54, no. 10, pp. 1065–1103, 2012.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[4] D. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *Proc. NIST Rich Transcript. Workshop*, 2004.

[5] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Univ. De Nice-Sophia Antipolis—UFR Sciences, Nice, France, Sep. 2005.

[6] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06's evaluation system," in *Proc. ICSLP*, 2006.

[7] T. H. Nguyen, H. Sun, S. Zhao, S. Z. K. Khine, H. D. Tran, T. L. N. Ma, B. Ma, E. S. Chng, and H. Li, "The IIR-NTU speaker diarization systems for rt 2009," in *Proc. RT'09, NIST Rich Transcript. Workshop*, 2009.

[8] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Comput. Speech Lang.*, vol. 20, no. 2, pp. 303–330, Jul. 2006.

[9] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 speaker diarization system: Enhancements in speaker modeling and cluster purification," in *Proc. ICASSP*, 2010, pp. 4958–4961.

[10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Ann. Appl. Statist.*, vol. 5, no. 2A, pp. 1020–1056, Jun. 2011.

[11] M. Johnson and A. Willsky, "The hierarchical dirichlet process hidden semi-markov model," in *Proc. Conf. Uncert. Artif. Intell.*, 2010.

[12] D. Blei and M. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–144, 2006.

[13] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 6, pp. 1059–1070, Dec. 2010.

[14] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proc. ICASSP*, 2008, pp. 4133–4136.

[15] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Proc. Interspeech*, 2011.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, Jul. 2010.

[17] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proc. IEEE Odyssey*, 2010.

[18] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Proc. Interspeech*, 2012.

[19] H. Ning, M. Liu, H. Tang, and T. Huang, "A spectral clustering approach to speaker diarization," in *Proc. ICSLP*, 2006.

[20] S. Shum, "Unsupervised methods for speaker diarization," M.S. thesis, Mass. Inst. of Technol., Cambridge, MA, USA, Jun. 2011.

[21] C. Vaquero, A. Ortega, and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," in *Proc. ICASSP*, 2011, pp. 4532–4535.

[22] D. Wang, R. Vogt, S. Sridharan, and D. Dean, "Cross likelihood ratio based speaker clustering using eigenvoice models," in *Proc. Interspeech*, 2011.

[23] J. Prazak and J. Silovsky, "Speaker diarization using PLDA-based speaker clustering," in *Proc. IDAACS*, 2011.

[24] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Proc. IEEE Odyssey*, 2012.

[25] T. Stafylakis, V. Katsouros, P. Kenny, and P. Dumouchel, "Mean shift algorithm for exponential families with applications to speaker clustering," in *Proc. IEEE Odyssey*, 2012.

[26] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[27] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey*, 2001.

[29] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[30] Diarization Error Rate (DER) Scoring Code. NIST, 2006 [Online]. Available: www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl

[31] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[32] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Univ. College London, London, U.K., May 2003.

[33] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS*, 2001.

[34] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. ASRU*, 2003.

[35] A. Martin and M. Przybocki, "Speaker recognition in a multi-speaker environment," in *Proc. Eurospeech*, 2001.

[36] E. Khan, J. Bronson, and K. Murphy, Variational Bayesian EM for Gaussian Mixture Models. 2008 [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/VBEMGMM/index.html

[37] M. J. Johnson, PYHSMM: A Python Library for Bayesian Inference in (HDP-)H(S)MMS. 2010 [Online]. Available: http://mattjj.github.com/pyhsmm/

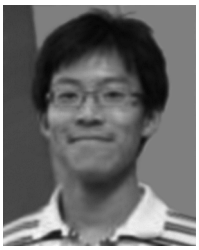
[38] D. Lin, E. Grimson, and J. Fisher, "Construction of dependent Dirichlet processes based on Poisson processes," in *Proc. NIPS*, 2010.

[39] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, vol. 26, pp. 189–206, 1984.

[40] H. Xu, C. Caramanis, and S. Mannor, "Principal component analysis with contaminated data: The high dimensional case," in *Proc. COLT*, 2010.

[41] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[42] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Proc. ICASSP*, 2011, pp. 4536–4539.



Stephen H. Shum is currently a Ph.D. student in Electrical Engineering and Computer Science (EECS) at the Massachusetts Institute of Technology (MIT). He received his B.S. in EECS at the University of California, Berkeley, in 2009 before joining the Spoken Language Systems Group at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), where he obtained his S.M. in 2011 and was awarded the William A. Martin Thesis Award for his work on speaker diarization. Although Stephen's research has primarily revolved around speaker recognition and diarization, his myriad interests also extend to semi-supervised learning, computational auditory scene analysis, and large-scale clustering of audio corpora.

Najim Dehak received his Engineering degree in Artificial Intelligence in 2003 from Universite des Sciences et de la Technologie d'Oran, Algeria, and his M.S. degree in Pattern Recognition and Artificial Intelligence Applications in 2004 from the Universite de Pierre et Marie Curie, Paris, France. He obtained his Ph.D. degree from Ecole de Technologie Superieure (ETS), Montreal in 2009. During his Ph.D. studies he was also with the Centre de recherche informatique de Montreal (CRIM), Canada. In the summer of 2008, he participated in the Johns Hopkins University, Center for Language and Speech Processing, Summer Workshop. During that time, he proposed a new system for speaker verification that uses factor analysis to extract speaker-specific features, thus paving the way for the development of the i-vector framework. Dr. Dehak is currently a research scientist in the Spoken Language Systems (SLS) Group at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He is also a member of IEEE Speech and Language Processing Technical Committee. His research interests are in machine learning approaches applied to speech processing and speaker modeling. The current focus of his research involves extending the concept of an i-vector representation into other audio classification problems, such as speaker diarization, language- and emotion-recognition.

Réda Dehak received his Ph.D. degree in signal and image processing from Ecole Nationale Supérieure des Télécommunication in 2002, his M.S degree in Signal, Image and Speech processing in 1998 from Institut National des Sciences Appliquées (INSA), Lyon, France and an Engineer degree in Computer Science in 1997 from Université des Sciences et de la Technologie d'Oran, Algeria. He is an assistant professor of computer science and member of the EPITA Research and Development Laboratory (LRDE). His research interests include speaker recognition, decision theory, pattern recognition and statistical learning. He is a member of the IEEE.

James R. Glass is a Senior Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) where he heads the Spoken Language Systems Group. He is also a Lecturer in the Harvard-MIT Division of Health Sciences and Technology. He received his B.Eng. in Electrical Engineering at Carleton University in Ottawa in 1982, and his S.M. and Ph.D. degrees in Electrical Engineering and Computer Science at MIT in 1985, and 1988, respectively. After starting in the Speech Communication group at the MIT Research Laboratory of Electronics, he has worked since 1989 at the Laboratory for Computer Science, and since 2003 at CSAIL. His primary research interests are in the area of speech communication and human-computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, supervised students, and published extensively in these areas. He is currently a Senior Member of the IEEE, an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and a member of the Editorial Board for Computer, Speech, and Language.