ISSN: 0975-5662, Dec, 2013 www.ijrct.org

A Survey on Speaker Diarization Approach for Audio and Video Content Retrieval

S. Sasikala M.Tech(MMT)*
Department of Computer Science and Engineering,
**K.S.R. College of Engineering*
ssksasi2@gmail.com

Dr. P. Balamurugan Ph.D*
Department of Computer Science and Engineering,
**K.S.R. College of Engineering*
pookumbala@gmail.com

Abstract:

Speaker diarization is the task of determining "who spoke when?" in an audio or video recording that contains an unknown amount of speech and also an unknown number of speakers. In the speaker diarization methods can be used to determine the speech part and non-speech part of the recordings. There are different approaches can be evaluated for speaker diarization. Accordingly, many important improvements in accuracy and robustness have been reported in journals and conferences in the area. It can be applicable on multiple distant microphones and single distant microphones. In this paper we review the current state-of-the-art, focusing on research developed. The analysis of the speaker diarization approaches can be delivered. Finally, we present an analysis of speaker diarization performance as reported through the NIST Rich Transcription evaluations on meeting data and identify important areas for future research.

Keywords: *Speaker diarization, Top down approach, Bottom up approach, Segmentation, Clustering, Features, Audio and video stream, NIST-RT evaluations, HMM, GMM.*

1 . INTRODUCTION

Speaker diarization has emerged as an increasingly important and dedicated domain of speech research. Whereas speaker and speech recognition involve, respectively, the recognition of a person's identity or the transcription of their speech, speaker diarization relates to the problem of determining 'who spoke when?' The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question, "Who spoke when?" While in speaker recognition, models are trained for a specific set of target speakers which are applied to an unknown test speaker for acceptance (target and test speaker match) or rejection (mismatch), in speaker diarization no prior knowledge about the identity or number of the speakers in the recording is given. Conceptually, a speaker diarization system therefore performs three tasks: First, discriminate between speech and non-speech regions (speech activity detection); second, detect speaker changes to segment the audio data; third, group the segmented regions together into speaker-homogeneous clusters.

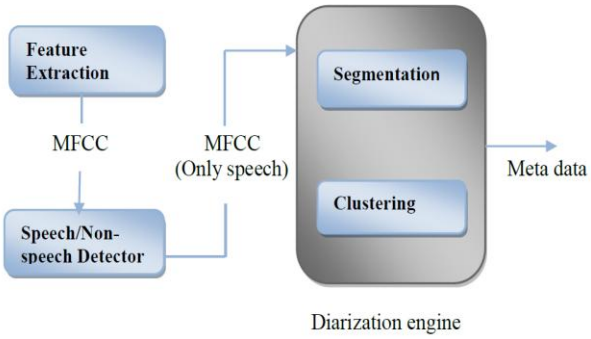

Fig. 1 Diagram illustrating the baseline ICSI Speaker Diarization Engine.

In ICSI speaker diarization engine has been developed. This engine extracts MFCC features from an audio track; it discriminates between speech and non-speech regions. It uses an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments and the grouping of these segments into speaker-homogeneous clusters in one step. Speech activity regions are determined using a state-of-the-art speech/non-speech detector. The detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible non-speech. The clustering approach based on agglomerative clustering approach. It is also called top down approach.

MFCC based features are also important to find the efficient audio features in the coefficients of each and every other term. [11] Introduces a diarization system with Hidden markov model (HMM) for each and every state defines the cluster rate that is speaker. Production probabilities can be calculated from one state to another state by using Gaussian

ISSN: 0975-5662, Dec, 2013 www.ijrct.org

Mixture Models (GMM). Each and every state takes 3 seconds to complete the process. After calculated the probabilities, realigns the result using viterbi alignment to purify the boundaries. This algorithm based on two categories such as, segmentation and clustering methods. It represents the three processing components namely, agglomerative clustering, Bayesian information criterion (BIC) and HMMs (Hidden markov models). It extends acoustic parameter such as Mel frequency Cepstral coefficients (MFCCs) to include features can be resulting from phone-classification multi-layer perceptrons (MLPs).

Multimodal Approach based on speaker diarization system such as speaker recognition or speaker identification methods. The architecture of this system can be decomposed into two main stages. First a reliable training set is created, in an unsupervised fashion, for each participant of the TV program being processed. This data is assembled by the association of visual and audio descriptors carefully selected in a clustering cascade. Then, Support Vector Machines are used for the classification of the speech data (of a given TV program).It involves three different steps that are, Feature extraction, Collecting Training Examples, and Hypothesized speaker's classification.

2. Approaches to Speaker Diarization

Speaker diarization can be defined as an optimization task on the space of speakers given the audio stream that is under NIST-RT evaluations. In the speaker diarization has two principal approaches, such as

- Bottom up approach
- Top down approach

The search procedure is initialized by over-segmenting or under-clustering the acoustic stream into a larger number of clusters than the assumed number of true speakers, and by training GMM models using the acoustic data in each cluster.

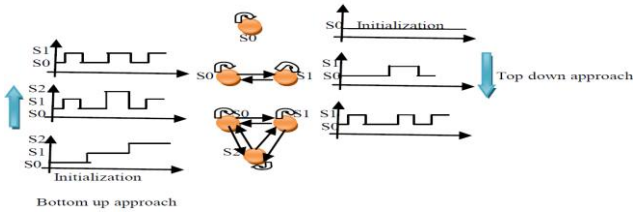

Fig. 2 Illustration of the bottom-up (left) and top-down (right) approaches to speaker diarization.

By using these approaches to constitute the segmentation and clustering methods. We first assume that non-speech segments have been removed from the acoustic stream and that features are extracted such that the remaining speech information is represented by a stream of acoustic features *O*. Letting *S* represent a speaker sequence and *G* a segmentation of the audio stream by *S*, then the task of speaker diarization can be formally defined as follows:

$$(\tilde{S}, \tilde{G}) = arg_{S,G} \max P(S, G \mid O) \quad ---- \quad (1)$$

Where $\tilde{S}$ and $\tilde{G}$ represent respectively the optimized speaker sequence and segmentation, i.e., who ($\tilde{S}$) spoke when($\tilde{G}$). In both approaches presented below the aim is to model each of the true speakers *N* with a single GMM. Speaker turns are represented by transitions between models thus forming an ergodic hidden Markov model (HMM) in which each state represents a speaker and where all states are fully connected.

2.1 Bottom-Up Approach

The bottom-up approach [20] is often referred to as agglomerative hierarchical clustering (AHC). The procedure is illustrated to the left of Fig.3 which shows how clustering begins with a larger speaker inventory (bottom) before similar clusters are merged to obtain a smaller, more optimal size (top). Only a single iteration is illustrated in Fig.2 and in this example the process stops when two clusters are obtained. The resulting diarization hypotheses are illustrated in the left column with the corresponding ergodic HMMs in the middle column.

Initialization produces three GMMs which are connected to form a three-state, ergodic HMM. Assume that the number of speakers is known approximately so that the bottom-up approach is initialized with more clusters than true speakers in order to avoid the risk of over-clustering. Here there are *N=2* true speakers and the bottom-up approach is initialized with three clusters. Using the initial HMM/GMMs the acoustic stream is re-segmented by Viterbi realignment before the models are refined according to the new segmentation. New models are re-estimated with expectation maximization (EM) procedure. The top-down approach to speaker diarization is less popular than its bottom-up counterpart but has however be shown to give competitive performance in NIST RT evaluations. The bottom-up approach is the most popular and has achieved general success in the NIST RT evaluations.

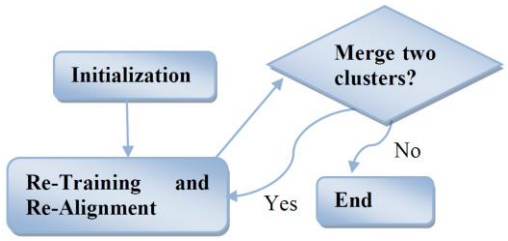

Fig. 3. Agglomerative clustering approach of the ICSI Speaker Diarization Engine

The initialization can be done by using k-clustering algorithm, where k denotes the number of speakers that are assumed in the recording. This algorithm can be processed by using these steps,

- Re-segmentation: To finding the optimal path of frames and models using viterbi alignment.
- Re-training: Given the new segmentation of the audio track, compute new Gaussian mixture models for each of the clusters.
- Cluster Merging: Given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing the BIC score (Bayesian information criterion).

2.2 Top down Approach

The top-down approach operates from a smaller speaker record to a larger speaker record and is a form of divisive hierarchical clustering (DHC). In the example shown to the right of Fig. 3, the approach starts with a single general speaker model and constructs an optimal speaker record by introducing new speakers one-by-one. As with the bottom-up approach, several iterations of Viterbi realignment and EM training are applied to process the model, until a stable segmentation is obtained.

New speakers are then added in the same way by the repeated splitting of existing models followed by several iterations of Viterbi realignment and EM training. The process continues until an optimal speaker record is obtained according to some stopping criteria, e.g., when there is no longer sufficient data with which to introduce a new speaker or when an upper limit on the size of the speaker record is reached.

Compared to the bottom-up approach, which reduces the number of models through cluster merging, the top-down approach increases the number of iteration through cluster splitting. Initialization is not good for top down approach and it can be shown in the fig.3. The two approaches thus have their own strengths and weaknesses and are therefore likely to exhibit different behavior and results. In the following we discuss some particular characteristics in further detail with the aim of better illuminating their potential merits such as, discrimination and purification, normalization and initialization.

3. Speaker Diarization Techniques

Given a probabilistic approach and stated assumptions the framework led to the two hierarchical clustering approaches to speaker diarization. The two systems are as general as is possible and are based on speaker diarization systems that have achieved state-of-the-art performance in the NIST RT'07 and RT'09 datasets evaluations can be estimated approximately in below Table I.

Dataset	FA	Miss	Total
RT'07	4.7	1.1	5.8
RT'09	7.2	1.8	9.0

TABLE I: Sad Performance on the RT'07 and RT'09 Datasets

3.1 Speech Activity Detection

SAD is a fundamental pre-processing step in all speaker diarization systems and aims to remove non-speech segments from the audio stream so that downstream speaker segmentation and clustering concentrates only on segments containing speech and also its better initialization for top down approach. ur SAD system follows standard noise suppression and is a simple model-based approach involving the alignment of the acoustic data to a two-state HMM in which the two states represent speech and non-speech data, respectively.

3.2 Bottom up Approach

Except for a novel progressive training approach to model initialization, which was proposed in and referred to as sequential EM, the first stage EM-based segmentation and clustering process is a conventional AHC approach as described. GMM speaker models contain four components but, in an otherwise standard AHC approach, they are initially trained using only a small fraction of the available data before several steps of re-estimation are performed with an increasing amount of data at each step. This process is repeated until all the data are used in the final training

cycle. New speaker models with 16 components are then estimated and used for the remaining merging steps.

3.3 Top-Down Approach

The top-down system is a DHC approach and is consistent with the general procedure. It is based on the evolutive hidden Markov model (E-HMM) that was originally proposed. The current system has evolved significantly from the original work and, with improvements to speaker modeling; the system was used for LIA-EURECOM's submission to the most recent NIST RT'09 evaluation.

3.4 Purification

Purification is a data filtering technique; the central idea is to remove noisy data so that models are trained on data that is indicative of the target class only and not of unwanted variation. Purification techniques have been extensively studied within bottom-up approaches to speaker diarization [28] but there is comparatively little work in the context of top-down approaches.

3.5 Combination

The bottom-up and top-down clustering strategies are likely to produce different diarization outputs and it is thus of interest to combine their outputs. We hypothesize that for both approaches, some models may reliably represent specific, individual speakers, whereas others may be relatively unreliable. In our work, to leverage the respective merits of both the bottom-up and top-down approaches, we treat the top-down output as a base segmentation and apply the bottom-up output to purify it. This set of reliable; matching clusters is denoted by Ξ. All unreliable or unmatched clusters are then compared to in order to identify additional reliable clusters, as follows:

$$\Xi \leftarrow C_m \quad \text{if} \quad l(C_m,\Xi) = max_k \; l\,(Ck,\Xi) C_k \notin \Xi \qquad (2), (3)$$

Specifically, for each cluster C_i contained in the top-down system output a cluster contained in the bottom-up system C_n is chosen as a matching cluster if

1) They share a sufficient proportion of frames and
2) Among all other clusters contained in the bottom-up system C_n is the closest to C_i , where the inter-cluster distance is measured in terms of ICR.

4. Components of Speaker Diarization

4.1 Feature Selection

Table II shows the initial list of candidate prosodic and long-term features. They are extracted using a library based on PRAAT. The features can be assigned to five different categories:

- Pitch,
- Energy,
- Formants,
- Harmonics-to-noise ratio, and
- Long-term average spectrum.

4.1.1 Pitch Features

Pitch features is based on pitch range available on each speakers. But vocal folds of the length and mass are unfair in the larynx. The main differences between the speakers are their gender and age with respect to pitch range. The vocal folds of post teenager men are longer and thicker with a lower modal frequency compared to women and children. Note that non-speech and unvoiced frames do not affect the results as pitch is marked as "undefined" in these cases.

- **Mean**: The average value.
- **Median**: The value of the 50th percentile. The median is generally less sensitive to outliers than the mean.
- **Min, max**: Instead of the actual minimal and maximal values, the 5th and 95th percentiles, respectively, are taken to avoid an otherwise large impact of outliers caused by artifacts.
- **Diff**: The difference between max and min as a measure of the local range.
- **Stdev**: The standard deviation as a measure of the variance.
- **Swoj**: The slope of the pitch curve ignoring octave jumps.

4.1.2 Energy Features

Compared to pitch, changes in loudness (or energy) are much less directly induced by anatomical characteristics.

Still, energy features are considered as potentially speaker discriminates and therefore used in the candidate list of features for this study. However, they are not hypothesized as getting a high rank. Energy features are based on an intensity contour with values in dB (SPL) i.e., dB relative to the human auditory threshold for 1 kHz. dB is used to logarithmic measurement of the energy level. The following statistics have been calculated using the above cases: min, max, diff (the difference between the former two), mean, and stdev.

4.1.3 Format Features

Formants are concentrations of acoustic energy around particular frequencies at roughly 1000-Hz intervals. However, formants also depend on the phonetic content. They occur only in voiced segments, i.e., vowels [a], nasals [n], liquids [l], voiced fricatives [v], and voiced stops [b].

4.1.4 Harmonics-to-Noise-Ratio Features

The harmonics- to-noise-ratio (HNR) quantifies the relative amount of additive noise in the voice signal. It is expressed in dB: if 99% of the energy of the signal is periodic and 1% is noise, then

$$HNR = 10*\log 10(99/1) = 20dB \quad ---- \quad (5)$$

An HNR of 0 dB corresponds to an equal amount of harmonic and non-harmonic energy.

4.1.5 Long-Term Average Spectrum (LTAS)

To obtain the long-term average spectrum, the spectral energy in 100-Hz-wide frequency bands is measured over a relatively large portion of speech (see Fig. 3). The standard deviation (stdev) was used as a measure of the variance. In addition, the slope of the curve (slope), the frequency associated with the lowest energy (fmin) and highest energy (fmax), and the peak heights (lph) were calculated. It would be computationally unfeasible to run diarization experiments with all subsets of this large set of prosodic features. Therefore, decided to perform a computational inexpensive pre-experiment to select the most promising features. A visualization of the feature histograms suggested that most features roughly follow a Gaussian distribution.

4.2 Speaker Clustering in Speaker-Diarization Systems

4.2.1 Speaker Clustering

The aim of speaker clustering in speaker-diarization systems is to associate or cluster together the segments from the same speaker. This clustering produces one cluster for each speaker, with all the segments from a given speaker in a single cluster. The dominant approach used in diarization systems is called hierarchical agglomerative clustering [6]; it consists of the following steps:

1. ***Initialization:*** each segment represents a single cluster;
2. ***Similarity measure:*** compute the pair-wise distances between each cluster;
3. ***Merging step:***
   - Merge the closest clusters together;
   - Update the distances of the remaining clusters to the new cluster;
4. ***Stopping criterion:*** iterate step 3 until some stopping criterion is met.

The presented agglomerative clustering approach is not the only possible solution for speaker clustering. This kind of approach is suitable in cases when all the audio data are available in advance. The speaker clustering involves on Bayesian Information Criterion (BIC) and Hidden Markov Model (HMM).

4.2.2 Spectral Clustering

Spectral clustering can handle very complex and unknown cluster shapes, and in this case the commonly used methods such as *K*-means and learning a mixture model using EM may fail. It relies on analyzing the Eigen-structure of a affinity matrix, rather than on estimating an explicit model of data distribution ([11], [15]). It is expected that it can also handle high dimensional audio data. We use a modification of the Ng-Jordan-Weiss (NJW) algorithm. For completeness of the text we first briefly review their algorithm.

As to this algorithm, there are still three open issues to be solved:

(1) Choice of metric, i.e. definition of d (s_i, s_j), and the fast algorithm to calculate it;
(2) Selection of the appropriate scale σ;
(3) Estimating automatically the number of clusters, i.e., *k*.

ISSN: 0975-5662, Dec, 2013 www.ijrct.org

4.3 Unsupervised Methods for Speaker Diarization

4.3.1Evaluation Protocol

Before specifying the methods, have to understand the evaluation methods. Setting the evaluation methods using NIST, so the Diarization Error Rate (DER) is the main 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.

The reference segmentation is a transcript of speech and speaker boundaries as given by the corpus. When measuring the evaluation performance, the evaluation code ignores intervals containing overlapped speech as well as errors of less than 250 ms in the locations of segment boundaries. Although overlapped speech intervals do not count in evaluating DER's.

4.3.2 Segmentation

In order to focus only on the speaker confusion portion of the Diarization Error Rate (DER) and not be mislead by mismatches between the reference speech/non-speech detector and our own (i.e., miss and false alarm errors), The use of previous work, 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 i-vector is extracted from each. Noting that this rather simple initial segmentation may result that contain speech from more than one speaker in each segments.

4.3.3 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 There exist many different ways to perform clustering such as, The Bayesian GMM and Its Variational Approximation , VBEM-GMM Clustering.

4.3.4 Final Pass Refinements

In this system, have to further refine the diarization output by extracting a single i-vector for each respective speaker using the (newly-defined) segmentation assignments. An 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.

4.4 Feature Extraction

As for the visual descriptors, we consider features characterizing the clothing of the TV show-participants, building upon the method initially proposed. Indeed, though the field size of the shots are varying (from long shots to close-ups), most of the time the person talking is seen on-screen. However, the use of a facial recognition system is rather difficult in our task due to the camera movements, the changing field-size and angles of shot, the changing postures of the filmed persons and the varying lighting conditions. The approach that we choose has the advantage that features relating to on-screen persons' clothing can be extracted even more robustly than the persons' facial features.

4.5 Collecting Training Examples

4.5.1 Shot detection

In the training set, the features can be given for identifying the faces for accurate results. In the shot detection, the human faces can be detected using speaker detection.

4.5.2 Lip activity detection

In the lip activity detection, the human faces can be extracted and the lip features is already given in the training set such as, talking can be identified by red rectangular boxes and yellow rectangular boxes is discarded for non-talking faces.

4.5.3 Video and audio clustering

Video Clustering: Using the cumulated HSV color histograms for the selected shots, a first (visual) grouping is performed. This one is a hierarchical agglomerative clustering. The distance used to measure the shot similarities is a distance computed on the

cumulated color histograms. It is defined as follows:

$$d_x^2 = \frac{1}{2}\sum_{i=1}^{b} \frac{(Hx_i - Hy_i)^2}{(Hx_i + Hy_i)} \quad \text{----} \quad (6)$$

Where Hx_i and Hy_i are the number of bins.

Audio Clustering: The audio clustering can be obtained by using SVM classification. The number of clusters can be sampled with the training sets.

4.6 SVM Audio Classification of Hypothesized Speakers

The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output. The classification task usually involves training and test sets which consist of data instances. The goal is now to process the remaining parts of the show (that were not selected during the data collection stage and not taken into account in the previous clustering cascade). Note that these remaining parts represent a higher fraction of the content, compared to the shots selected in the previous stage.

Practically, it uses one-vs-one SVM classifiers, meaning that for the hypothesized speakers *N* obtained earlier, bi-class classifiers are trained. Since the training sets are potentially imbalanced, it uses a different *C* value for positive and negative training examples, which will refer to as *C*+ and *C*-respectively.

5. CONCLUSION AND FUTUREWORK

In the speaker diarization systems have evaluated different approaches for different talk shows, conference meetings, debate, political debates and so on. By NIST-RT evaluations was very helpful for the speaker diarization performance. This article provides an overview of the current state-of-the-art in speaker diarization systems and underlines several challenges that need to be addressed in future years. Speaker diarization is one of the fundamental problems underlying virtually any task that involves acoustics and the presence of more than one person. 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. Based on the above researches, SVM classification techniques gives better improvement of speaker diarization approaches, so the future work may be using SVM classification for improving the audio-visual content. The evaluation between our approach and modern approaches designate that our approach is strong and efficient

6. REFERENCES

1. D. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP’05)*, Mar. 2005, vol. 5, pp. 953–956.
2. B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, “Speaker role recognition to help spontaneous conversational speech detection,” in *Proc. ACM Workshop Searching for Spontaneous Conversational Speech*, Firenze, Italy, Oct. 2010.
3. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, “A Practical Guide to Support Vector Classification, April 15, 2010.
4. D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, “A Privacy-sensitive approach to modeling multi-person conversations.,” *Proc. IJCAI*, 2007.
5. Dielmann, “Unsupervised detection of multimodal clusters in edited recordings,” in *Proc. Multimedia Signal Processing*, Saint-Malo, France, Oct. 2010.
6. Do, M. N. (2003). Fast Approximation of Kullback-Lebler Distance for Dependence Trees and Hidden Markov Models. *Signal Processing Letters*, Vol. 10, (2003), pp. 115-118.
7. F. Vallet, S. Essid, J. Carrive, and G. Richard, “Robust visual features for the multimodal identification of unregistered speakers,” in *Proc. Int. Conf. Image Processing*, Hong Kong, China, Oct. 2010.
8. Félicien Vallet, Slim Essid, and Jean Carrive, ” A Multimodal Approach to Speaker Diarization on TV Talk-Shows”, *IEEE Trans. Multimedia*, vol. 15, no.3, pp. 509-520 , APRIL 2013.
9. G. Friedland, H. Hung, and C. Yeo, “ Multimodal Speaker diarization of real world meetings using compressed domain video features,” in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009.
10. G. Richard, M. Ramona, and S. Essid, “Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams,” in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, Apr. 2007.
11. J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Proc. IEEE Autom. Speech Recognition Understand. Workshop*, 2003, pp. 411–416.
12. M. Bendris, D. Charlet, and G. Chollet, “Lip activity detection for talking faces classification in TV-content,” in Proc. Int. Conf.Machine Vision, Hong Kong, China, Dec. 2010.

13. S. Bozonnet, F. Vallet, N. Evans, S. Essid, G. Richard, and J. Carrive, "A multimodal approach to initialization for top-down speaker diarization of television shows," in *Proc. European Signal Processing Conf.*, Aalborg, Denmark, Aug. 2010.
14. 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'10*, Dallas, TX, Mar. 14–19, 2010, pp. 4958–4961.
15. 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. Special Iss. New Frontiers in Rich Transcript*., 2012.

