An Open-source State-of-the-art Toolbox for Broadcast News Diarization

Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, Sylvain Meignier

An Open-source State-of-the-art Toolbox for Broadcast News Diarization

Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, Sylvain Meignier

LUNAM Université, LIUM, Le Mans, France

first.lastname@lium.univ-lemans.fr

Abstract

This paper presents the LIUM open-source speaker diarization toolbox, mostly dedicated to broadcast news. This tool includes both Hierarchical Agglomerative Clustering using well-known measures such as BIC and CLR, and the new ILP clustering algorithm using i-vectors. Diarization systems are tested on the French evaluation data from ESTER, ETAPE and REPERE campaigns.

Index Terms: speaker diarization, broadcast news, open-source

1. Introduction

Speaker diarization, the "who spoke when" task, consists in annotating recordings with labels representing the speakers. This task is performed without any prior information: neither the number of speakers, nor their identities, nor samples of their voices are available.

Since 2004, the state-of-the-art system for broadcast news speaker diarization is composed of 5 steps. First, music and jingle regions are removed using a Viterbi decoding. Next, an acoustic segmentation followed by a Hierarchical Agglomerative Clustering (HAC) splits and then groups the signal into homogeneous parts according to speakers and background. In this step, each segment or cluster is modeled by a Gaussian distribution with a full covariance matrix and the Bayesian Information Criterion (BIC) is employed both as similarity measure and as stop criterion. Then, a Gaussian Mixture Model (GMM) is trained for each cluster via the Expectation-Maximization (EM) algorithm. The signal is then re-segmented through a Viterbi decoding. The system finally performs another HAC, using the Cross-Likelihood Ratio (CLR) measure and GMMs trained with the Maximum *A Posteriori* algorithm (MAP). This kind of system obtained the best results at NIST RT'04 fall [1], ESTER [2, 3], ETAPE [4] and, more recently, REPERE [5] evaluation campaigns.

LIUM_SpkDiarization[1] is an application dedicated to processing radio and TV shows. It has been developed to provide a ready and easy to use tool for the multimedia community. *LIUM_SpkDiarization* is also a toolbox which allows the development of new diarization systems, either by creating scripts from basic tools (segmentation, classification, Viterbi decoder, etc.), or by adding new functionality directly in the source code. *LIUM_SpkDiarization* was not developed from scratch. It has evolved from a previous speaker segmentation tool, *mClust*, developed in C++ by LIUM for the French ESTER evaluation campaigns in 2005 and 2008. As a Java-based descendant of *mClust*, *LIUM_SpkDiarization* has adopted an internal architecture very close to that of its ancestor. The first public version of *LIUM_SpkDiarization* was made available at the 2010 Sphinx Workshop [6]. This paper presents a review of the tools included in the toolkit, and describes the latest improvements.

In the next section, we present other diarization tools. Then, we describe the choices made for this tool before describing the diarization system for broadcast news, in section 3. The single-show speaker diarization system is described in section 4, and the cross-show speaker diarization is reported in section 5. In section 6, we describe experiments performed on various corpora and give results. Finally, section 7 describes a video shot boundary detector to illustrate how this toolbox could be employed in other fields.

2. Other diarization tools

Several toolkits distributed under open-source licenses are available on the web. One of the oldest is the *CMU Segmentation* tool which was released in 1997 [7]. It was developed during the former NIST broadcast news evaluation campaign to address specifically the task of diarization for automatic speech recognition.

AudioSeg [8], under the GPL license, is a toolbox developed by IRISA during the ESTER campaign in 2005. It includes an audio activity detector, BIC/GLR or KL2 segmentation and clustering tools, as well as a Viterbi decoder. Note that CLR-based clustering is not available.

The speaker recognition library *ALIZÉ* [9] also includes speaker diarization tools. Diarization is based on the E-HMM method [10] in which segmentation and clustering are done iteratively and jointly. Performance is better when dealing with meetings and phone conversations rather than broadcast news.

The speech recognition toolkit *SHoUT* includes a speech/non-speech detector and a diarization tool. This tool seems to be well adapted for recordings of meetings as shown by the results reported in [11].

Recently, IDIAP published *DiarTK* [12] where clustering and segmentation are based on the information bottleneck principle. It was developed specifically for meetings recorded using multiple distant microphones or microphone arrays.

3. Guidelines of the toolkit

LIUM_SpkDiarization is developed in Java to minimize dependency problems with the various operating systems and libraries. This tool is distributed as one self-contained JAR archive, which can be run directly, with no need of additional third-party packages. Indeed, all the required packages are included in the precompiled version, as well as models (UBM, gender or speech/non-speech models).

LIUM_SpkDiarization is a simple program for those who only need to perform speaker diarization for their own applications (speech recognition, speaker recognition, multimedia in-

[1]This research was partly supported by ANR (French National Research Agency) under contract number ANR-2010-CORD-101-01 (SODA project).

dexing, etc.). The execution of the JAR archive calls upon the speaker diarization method dedicated to broadcast news recordings. The acoustic parameters (MFCC) are computed directly by the program using Sphinx4 (one of the included third-party packages).

LIUM_SpkDiarization is also designed to carry out research. The toolkit is composed of elementary programs, such as segment and cluster generator, decoder, and model trainers. Fitting those elementary tools together in a shell script is an easy way to develop a specific diarization system.

This toolbox can employ alien features. By default, the tools read Sphinx MFCC files; however, it also reads HTK, SPro files, as well as text files where each line corresponds to a frame. Regarding speaker models, the toolkit has its own binary file format but is also able to read ALIZÉ files.

LIUM_SpkDiarization allows the management of a speaker diarization which works with clusters of segments. A segment is an object defined by the name of the recording it belongs to, its start time in the recording, and its duration. These three values identify a unique region of signal in a set of audio recordings. The elementary programs that compose the toolkit are able to work with a set of audio recordings. This way, models can be learnt from parts of recordings without having to cut them, making cross-show diarization systems easy to develop (see section 5).

4. Single-show diarization system

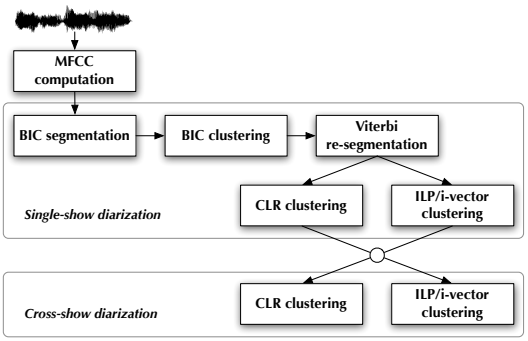

Figure 1: Diarization steps

The proposed diarization system was inspired by the system [13] which won the RT'04 fall evaluation as well as the ESTER evaluation in 2005. It was developed for transcription and diarization tasks, and with the goal of minimizing both Word Error Rate and speaker error rate (*i.e.* Diarization Error Rate – DER).

Automatic transcription requires accurate segment boundaries. Segment boundaries have to be set within non-informative zones such as filler words. Indeed, having a word cut by a boundary disturbs the language model and increases the Word Error Rate; whereas the main objective of speaker diarization is to produce the minimum number of pure clusters, each of which corresponds to only one single speaker.

Non-speech segments must be rejected to save computation time in both tasks: in transcription, non-speech segments generate insertion of words and, in diarization, non-speech segments make the speaker models less accurate.

4.1. Diarization for speech transcription

This system is widely described in [6]. Only an outline is given below:

1. The acoustic frames are composed of 13 MFCCs with coefficient C_0, and feature normalization is not applied. In the two last steps, the acoustic frames are composed of 12 MFCCs (C_0 is removed) as well as the first-order derivatives of those coefficients.
2. The speaker segmentation is composed of two passes. The first detects the instantaneous change points using Generalized Likelihood Ratio (GLR) distance. The second pass over the signal uses BIC distances between speakers in order to fuse consecutive segments that are found to correspond to the same speaker.
3. The Hierarchical Agglomerative Clustering merges the two closest clusters at each iteration until the best BIC distance is positive. In this step and in the previous one, speakers are modeled by a Gaussian distribution with a full covariance matrix.
4. A Viterbi decoding is performed to generate a new segmentation using GMMs as speaker models. C_0 is removed from the features and first-order derivatives of the coefficients are added.
5. A segmentation into speech/non-speech is obtained using a Viterbi decoding in order to remove music and jingle regions. The initial segmentation is less accurate if this detection step is done before the speaker segmentation, like most of the diarization systems do. The main reason is that speaker segmentation cannot make correct decisions on the 2.5 seconds at the beginning and at the end of a segment. This problem is exacerbated when the speech/non-speech segmentation is done first because it generates many segments to cut.
6. The gender and bandwidth of clusters are detected using a GMM over normalized frames as in CLR-like clustering described below. The segment boundaries produced are not perfect: for example, some of them fall within words. In order to avoid this situation, boundaries are moved to low energy regions. Long segments are also cut in order to yield segments shorter than 20 seconds.

4.2. Diarization optimized for speaker clustering

Step 6 described above was not required in order to perform an accurate speaker diarization, because only gender and bandwidth information, as well as boundary location, are needed for speech transcription. In the current case, another clustering step is performed instead.

4.2.1. CLR-like Clustering

The second Hierarchical Agglomerative Clustering merges the two closest clusters at each iteration until the similarity between the two candidate clusters is positive. Most systems use the CLR measure to estimate similarity between clusters [14], but it has been shown that normalized CLR measure [15] gives better results in most cases. The difference between CLR and normalized CLR resides only in the denominators, which are the UBM log-likelihood scores and the log-likelihood scores of the clusters respectively. In this step, the first order derivative is added to the 12 MFCCs (C_0 is removed) and the features are normalized: short-term windowed mean and variance are computed

to normalize the frame, and a feature warping normalization is applied. Only the means of the Universal Background Model (UBM) are adapted for each cluster in order to obtain the model for its speaker.

4.2.2. ILP/i-vectors clustering

A new clustering algorithm in which the problem is solved as an Integer Linear Programming (ILP) problem has been proposed [16]. Experiments were carried out using i-vectors to model and measure the similarity between clusters. This method is now available in the *LIUM_SpkDiarization* toolbox.

The i-vectors is a state-of-the-art method in the field of Speaker Verification [17]. The acoustic data of a speaker are compacted into a low-dimension vector, which only retains the relevant information about the speaker. This approach was first adapted to 2-speaker diarization using the *k-means* algorithm to find utterances of the two speakers within a phone call recording [18].

In broadcast news, the number of speakers is unknown *a priori* . The proposed method replaces the standard CLR-like clustering step. According to an initial speaker segmentation, an i-vector is extracted from each cluster using 19 MFCCs parameters completed with energy, their first and second order derivatives, and a 1024 GMM-UBM. The N resulting i-vectors are then normalized in an iterative process [19]. The clustering problem consists in jointly minimizing the number K of cluster centers as well as the dispersion of i-vectors within each cluster. The value $K \in \{1, \ldots, N\}$ is to be automatically determined.

This clustering problem is expressed as an ILP problem, where the objective solving function (eq. 1) is minimized subject to constraints:

Minimize

$$\sum_{k=1}^{N} x_{k,k} + \frac{1}{D} \sum_{k=1}^{N} \sum_{j=1}^{N} d(k,j) x_{k,j} \qquad (1)$$

Subject to

$$x_{k,j} \in \{0,1\} \qquad \forall k, \forall j \qquad (1.2)$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad \forall j \qquad (1.3)$$

$$d(k,j) x_{k,j} \leq \delta \qquad \forall k, \forall j \qquad (1.4)$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad \forall j \qquad (1.5)$$

Where $x_{k,k}$ (eq. 1) is a binary variable equal to 1 when the i-vector k is a center. The number of centers K is implicitly included in equation 1, indeed $K = \sum_{k=1}^{N} x_{k,k}$. The distance $d(k,j)$ is computed using the *Mahalanobis* distance between i-vectors k and j. D is a normalization factor equal to the longest distance $d(k,j)$ for all k and j. The binary variable $x_{k,j}$ is equal to 1 when the i-vector j is assigned to the center k. Each i-vector j will be associated with a single center k (eq. 1.3). The i-vector j associated with the center k (*i.e.* $x_{k,j} = 1$) must have a distance $d(k,j)$ shorter than a threshold δ empirically determined (eq. 1.4). Equation 1.5 ensures that the cluster k is selected if an i-vector is assigned to cluster k.

5. Cross-show diarization system

The recently introduced cross-show speaker diarization [20, 21, 22] aims to expand the diarization task to a broader context, where speakers appearing in different recordings of the same show (the *cross-show speakers*) will always be identified in the same way in every recording. Each show from a collection is first individually processed with the single-show speaker diarization system, as described in section 4. Then it is processed collectively using a CLR or ILP clustering. Experiments showed that ILP clustering provides a better speed/accuracy trade-off [22].

6. Single- and cross-show diarization evaluation

6.1. Data

Evaluation has been performed on three different corpora:

- The test corpus of the ESTER 2[2] evaluation campaign is composed of 26 broadcast news recordings coming from seven French radio stations. Most of the corpus contains prepared speech. 15 hours of the shows are fully annotated.
- The test corpus of the ETAPE[3] evaluation campaign is composed of 15 TV shows recorded from two French TV channels and one French radio station. Most of the corpus contains spontaneous speech or very spontaneous speech. 7 hours of the shows are fully annotated.
- The test corpus of the REPERE 2012[4] evaluation campaign is composed of 3 hours of data drawn from 28 TV shows recorded from French TV channels BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, from 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

6.2. Evaluation metrics

The Diarization Error Rate (DER) is the metric used to measure performance. DER was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker using the best match between references and hypothesis speaker labels. The scoring tool we used was developed by the LNE[5] as part of the ETAPE and the REPERE campaigns.

The LNE evaluation tool computes two different error rates. The *single-show DER* is computed considering each show independently. In this context, the overall DER corresponds to the mean of individual DERs (one per show), weighted by the duration of each show. The *cross-show DER* is computed over all the shows, taking into account multiple appearances of the same speaker in several shows. In order to assess the cross-show diarization task, the same label must necessarily identify speakers appearing in several shows.

6.3. Results and comments

Single- and cross-show speaker DER on the three test corpora are reported in Table 2. Both CLR and ILP clustering methods were tested. The cross-show CLR clustering was performed using the single-show CLR segmentations, and the cross-show ILP clustering with the single-show ILP segmentations.

[2] `http://catalog.elra.info`, reference ELRA-S0338
[3] `http://www.afcp-parole.org/etape-en.html`
[4] `http://www.defi-repere.fr`
[5] The French National Laboratory of Metrology and Testing

Tools	Feature	Description
MFCC	format	Sphinx, HTK, SPRO, text
	normalization	mean (+variance) by show, cluster, segment or in a sliding window, feature mapping
	derivative	first and/or second order with or without energy, Sphinx or SPro formula
	frames discarding	energy threshold or bi-Gaussian on energy
Segmentation	distance (set 1)	GLR or BIC using Gaussian (full or diagonal)
	distance (set 2)	KL2, GD[13] or ICR[23] using diagonal Gaussian
	change points	recursive or local minimum search form left to right
HAC	distance (set 1)	KL2 using diagonal Gaussian or GMM [24]
	distance (set 2)	CLR, normalized CLR[15] or T-test [25] using GMM
	distance (set 3)	BIC using Gaussian (full or diagonal)
	BIC penalty factor	global $log(N)$, local $log(n_i + n_j)$, square root BIC [26]
Other clustering	E-HMM	2-speakers E-HMM segmentation and classification with MAP adaptation [10]
	Meeting	based on IDIAP/ICSI meeting method (experimental) [27]
	ILPILP/i-vector	model computed using ALIZÉ [9]
Training EM	iteration control	minimum and maximum iterations
		minimum gain of likelihood per iteration
	variance control	ceiling and flooring
Training MAP	algorithm	standard, linear, Variable Prior MAP [28]
	adaptation of	weight and/or mean and/or variance
Viterbi	model	pre-computed GMM only
	penalty	for each HMM state, set the loop and exit penalties
	duration	n states minimum (default $n = 1$), change only to multiple-of-n states
Data	model	8 GMMs for speech/non-speech
		3 GMMs for silence
		512-components GMM
		4x128-components GMM for gender / bandwidth detection
		model for i-vector

Table 1: Tools and features of the LIUM_SpkDiarization v8.0

Because the segmentation is unchanged during the clustering steps, *False Alarms* (FA) and *Missed Detections* (MISS) obtained on a corpus remains the same, regardless of the clustering method or the evaluation metric. The values are as follows: 1.58% FA and 0.98% MISS on the ESTER 2 test corpus; 4.10% FA and 4.24% MISS on the ETAPE test corpus; 1.10% FA and 3.83% MISS on the REPERE 2012 test corpus.

It is important to note that the systems were not specifically tuned according to the test corpora, and better DER could be obtained by doing so. Single- and cross-show CLR thresholds were set to 0.35 and 0.80 respectively. Single- and cross-show ILP distance thresholds were set to 100 and 60 respectively.

Type	Corpus	Single-show DER	Cross-show DER
CLR	ESTER 2	11.27 %	20.43 %
	ETAPE	21.57 %	27.79 %
	REPERE	17.19 %	23.95 %
ILP	ESTER 2	8.35 %	17.51 %
	ETAPE	24.49 %	26.31 %
	REPERE	15.46 %	19.59 %

Table 2: Single- and cross-show DER on the ESTER 2, ETAPE and REPERE 2012 test corpora.

7. Extension to computer vision tasks

Recent studies show that there is quite a large similarity between audio and video tasks, especially for solving several identification and segmentation problems. In [29], the authors show that successful speaker identification techniques are also very good for face identification. Furthermore, in [30], the authors show that speaker segmentation techniques using the generalized likelihood ratio (GLR) and the Bayesian information criterion (BIC) can also be used for other segmentation tasks such as video shot boundary detection or TV program boundary detection. In this work, we give the example of shot boundary detection to show how easily *LIUM_SpkDiarization* can be extended. Shot boundary detection (SBD) is a well-known segmentation process. It aims at breaking down the massive volume of video into smaller chunks. Quite a lot of approaches have been proposed in the literature in the last two decades [31]. The main ideas behind their assumption is that 1) shots, similarly to audio turns, are homogeneous segments, and 2) color features similarly to MFCC features, can be modeled by Gaussian distributions.

Because of this similarity, the implementation of SBD in *LIUM_SpkDiarization* is straightforward. Although the visual feature extractor module is missing, *LIUM_SpkDiarization* has a good option that allows the user to use external features, and in various file formats. This option makes the integration of new tasks very simple. Consequently, results obtained by our SBD implementation are equivalent to the best state-of-the-art systems (F-measure $> 95\%$ on TV Broadcast news and debates).

8. Conclusions

This paper presents a diarization toolkit mostly dedicated to broadcast news recordings. Developed by the LIUM, this toolkit is published under the GPL license. When used for broadcast news diarization, it represents the state-of-the-art in terms of performance; in addition, only minor work is required in order to reuse its components for other tasks. The *LIUM_SpkDiarization* toolbox, as well as its documentation and some examples, is available at: `http://www-lium.univ-lemans.fr/en/content/liumspkdiarization`.

9. References

[1] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," August 2004. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf

[2] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 05)*, Lisbon, Portugal, September 2005, pp. 1149–1152.

[3] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proceedings of Interspeech*, September 2009.

[4] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC - Eighth international conference on Language Resources and Evaluation*, Turkey, 2012.

[5] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE corpus : a multimodal corpus for person recognition," in *LREC - Eighth international conference on Language Resources and Evaluation*, 2012.

[6] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, Texas (USA), March 2010.

[7] M. Siegler, U. U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *the DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.

[8] G. Gravier, M. Betser, and M. Ben, *audioseg : Audio Segmentation Toolkit, release 1.2.*, IRISA, january 2010.

[9] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Odyssey: the Speaker and Language Recognition Workshop*, 2008.

[10] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

[11] M. Huijbregts, "Segmentation, diarization and speech transcription: Surprise data unraveled," Ph.D. dissertation, Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, 2008.

[12] D. Vijayasenan and F. Valente, "DiarTk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proceedings of Interspeech*, Portland, Oregon (USA), 2012.

[13] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multi-stage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[14] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.

[15] V.-B. Le, O. Mella, and D. Fohr, "Speaker diarization using normalized cross-likelihood ratio," in *Proceedings of Interspeech*, 2007.

[16] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Odyssey Workshop*, Singapore, 2012.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," in *Proceedings of IEEE TASLP*, vol. 19, 2011, pp. 788–798.

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

[19] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[20] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing multi-stage approaches for cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[21] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[22] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "I-vectors and ILP clustering adapted to cross-show speaker diarization," in *Proceedings of Interspeech*, Portland, Oregon (USA), 2012.

[23] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Proceedings of Interspeech*, 2007, pp. 1853–1856.

[24] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between GMMs," in *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, Jeju, Korea, 2004.

[25] T. H. Nguyen, E. S. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in *Interspeech 2008*, September 2008.

[26] T. Stafylakis, G. Tzimiropoulos, V. Katsouros, and G. Carayannis, "A new penalty term for the BIC with respect to speaker diarization," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4978 –4981.

[27] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[28] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proceedings of Interspeech*, 2005, pp. 2437–2440.

[29] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian Mixture Models," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 553 – 562, 2012.

[30] E. Khoury, C. Sénac, and P. Joly, "Unsupervised segmentation methods of TV contents," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, March 2010.

[31] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.