# Speaker Change Detection using Binary Key Modelling with Contextual Information

Jose Patino, Héctor Delgado, and Nicholas Evans

Department of Digital Security
EURECOM
Sophia Antipolis, France
{patino,delgado,evans}@eurecom.fr

**Abstract.** Speaker change detection can be of benefit to a number of different speech processing tasks such as speaker diarization, recognition and detection. Current solutions rely either on highly localized data or on training with large quantities of background data. While efficient, the former tend to over-segment. While more stable, the latter are less efficient and need adaptation to mis-matching data. Building on previous work in speaker recognition and diarization, this paper reports a new binary key (BK) modelling approach to speaker change detection which aims to strike a balance between efficiency and segmentation accuracy. The BK approach benefits from training using a controllable degree of contextual data, rather than relying on external background data, and is efficient in terms of computation and speaker discrimination. Experiments on a subset of the standard ETAPE database show that the new approach outperforms the current state-of-the-art methods for speaker change detection and gives an average relative improvement in segment coverage and purity of 18.71% and 4.51% respectively.

**Keywords:** speaker change detection, binary keys, speaker diarization, speaker recognition

## 1 Introduction and related work

Speaker change detection (SCD), also known as speaker turn detection and more simply speaker segmentation, aims to segment an audio stream into speaker-homogeneous segments. SCD is often a critical pre-processing step or enabling technology before other tasks such as speaker recognition, detection or diarization.

The literature shows two general approaches. On the one hand, metric-based approaches aim to determine speaker-change points by computing distances between two adjacent, sliding windows. Peaks in the resulting distance curve are thresholded in order to identify speaker changes. Bayesian information criterion (BIC) [8] and Gaussian divergence (GD) [4] are the most popular metric-based approaches. On the other hand, model-based approaches generally use off-line training using potentially large quantities of external data. Example model-based approaches utilise Gaussian mixture models (GMMs) [20], universal background

models (UBMs) [24], or more recent techniques such as those based on the i-vector paradigm [25, 21] or deep learning [17, 6, 23, 18].

Despite significant research effort, SCD remains challenging, with high error rates being common, particularly for short speaker turns. Since they use only small quantities of data within the local context, metric-based approaches are more efficient and domain-independent, though they tend to produce a substantial number of false alarms. This over-segmentation stems from the intra-speaker variability in short speech segments. Model-based approaches, while more stable than purely metric-based approaches, depend on external training data and hence may not generalise well in the face of out-of-domain data.

The work reported in this paper has sought to combine the merits of metric- and model-based approaches. The use of external data is avoided in order to promote domain-independence. Instead, the approach to SCD reported here uses variable quantities of contextual information for modelling, i.e. intervals of the audio recording itself. These intervals range from the whole recording to shorter intervals surrounding a hypothesized speaker change point.

The novelty of the approach lies in the use of an efficient and discriminative approach to modelling using binary keys (BKs) which have been reported previously in the context of speaker recognition [1, 5], emotion recognition [3, 19], speech activity detection [15] and, more extensively, speaker diarization [15, 2, 12, 11, 14, 13, 22]. In all of this work, segmentation consists in a straightforward partition of the audio stream into what are consequently non-heterogeneous speaker segments. In this case, speaker segmentation is only done implicitly at best; none of this work has investigated the discriminability of the BK approach for the task of explicit SCD. The novel contribution of this paper includes two BK-based approaches to explicit SCD. They support the flexible use of contextual information and are both shown to outperform a state-of-the-art baseline approach to SCD based on the Bayesian information criterion (BIC).

The remainder of the paper is organised as follows. Section 2 describes binary key modelling. Its application to speaker change detection is the subject of Section 3. Section 4 describes the experimental setup including databases, system configuration and evaluation metrics. Section 5 reports experimental results and discussion. Conclusions are presented in Section 6.

## 2 Binary key modelling

This section presents an overview of the binary key modelling technique. The material is based upon original work in speaker diarization [2] and recent enhancements introduced in [22].

### 2.1 Binary key background model (KBM)

Binary key representations of acoustic features are obtained using a binary key background model (KBM). The KBM plays a similar role to the conventional

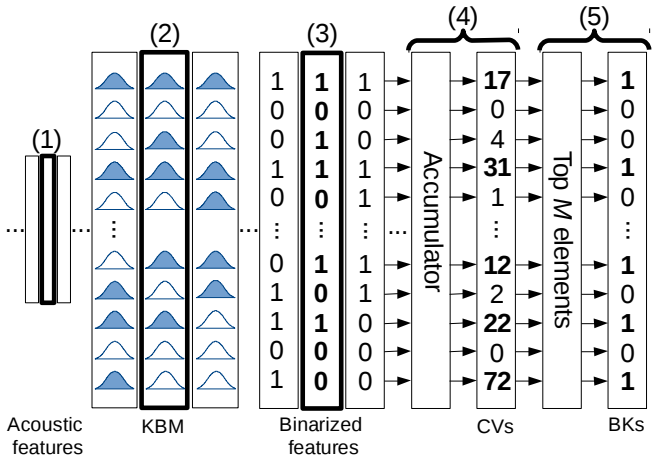


**Fig. 1.** An illustration of the BK extraction procedure based upon the comparison of acoustic features and the KBM

UBM. Just like a UBM, the KBM can be estimated using either external data [1] or the test data itself [2].

The KBM training procedure involves the selection of discriminative Gaussians from a large pool of candidates. The pool is learned from standard acoustic features extracted with a conventional frame-blocking approach and spectral analysis. The selection process uses a single-linkage criterion and a cosine distance metric to select the most discriminant and complementary Gaussians. This approach favours the selection of dissimilar candidates thereby resulting in a set of Gaussians with broad coverage of the acoustic space; closely related, redundant candidates (which likely stem from the over-sampling of homogeneous audio segments) are eliminated. The process is applied to select an arbitrary number of $N$ Gaussian candidates which then constitute the KBM. As in [22], the size of the KBM is expressed not in terms of a fixed number of components, but is defined adaptively according to the quantity of available data. The resulting KBM size is hence defined as a percentage $\alpha$ of the number of Gaussians in the original pool. The KBM is hence a decimated version of a conventional UBM where Gaussian components are selected so as to be representative of the full acoustic space in terms of coverage rather than density. Full details of the KBM training approach can be found in [13].

### 2.2 Feature binarisation

Binarised features are obtained from the comparison of conventional acoustic features to the KBM. The process is illustrated in Figure 1. A sequence of $n_f$ acoustic features is transformed into a binary key (BK) whose dimension $N$ is dictated by the number of components in the KBM. For each acoustic feature vector in the input sequence (labelled 1 in Figure 1), the likelihood given each

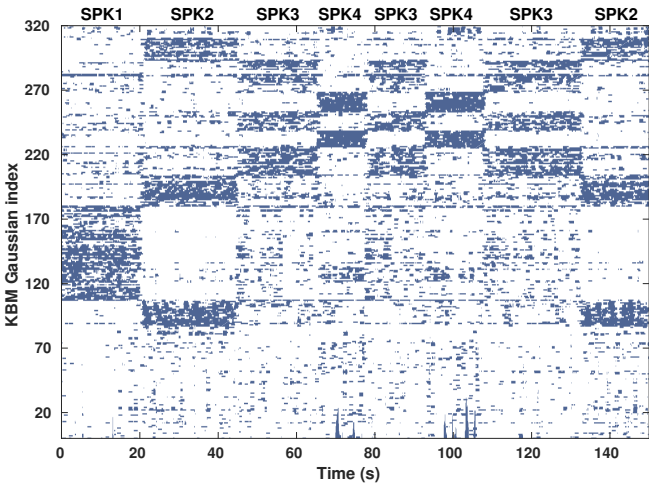


**Fig. 2.** A matrix of BKs from an arbitrary 2.5-minute speech fragment from the ETAPE database. Each column of the matrix is an individual BK with $N$=320 elements extracted according to the procedure illustrated in Figure 1. Distinguishable BK patterns indicate distinct speakers whereas differences bewteen them indicate speaker change points

of the $N$ KBM components is computed and stored in a vector which is sorted by Gaussian index. The top $N_G$ Gaussians defined as those with the $N_G$ highest likelihoods (2 - illustrated in solid blue) are then selected and used to create binarised versions of the acoustic features (3).

This process is repeated for each frame of acoustic features thereby resulting in a binary matrix of $n_f \times N$, each column of which has $N_G$ values equal to binary 1. A row-wise addition of this matrix is then used to determine a single cumulative vector (CV) which reflects the number of times each Gaussian in the KBM was selected as a top-Gaussian (4). The final BK is obtained from the $M$ positions with highest values in the CV (5). Corresponding elements in the BK are set to binary 1 whereas others are set to 0. The BK provides a sparse, fixed length representation of a speech segment based on similarities to the acoustic space modelled by the KBM. Full details of the feature binarisation approach are also available in [13].

### 2.3 An illustrative example

By way of illustrating the speaker-discriminability of the BK approach, Figure 2 depicts a sequence of BKs extracted following the procedure described in Section 2.2 from an arbitrary speech fragment in the order of 2.5 minutes duration. Each column of the matrix is a BK computed from a 1s window with a 0.1s shift using a KBM of size $N = 320$. Speaker labels towards the top of the plot indicate the speaker which is active during each apparent segment. The vertical axis indicates the sorted KBM Gaussian indexes whereas the horizontal axis

indicates time. The intra-speaker consistency of BKs is immediately evident, as are the inter-speaker differences which indicate speaker changes or turns. The apparent diagonal component towards the upper half of the figure stems from the sequential, temporal nature with which Gaussian candidates are added to the KBM pool.

## 3 BK-based speaker change detection

This section describes the application of BK modelling to speaker change detection. Two such approaches are proposed.

### 3.1 KBM estimation

The KBM can be learned using either one of the two approaches illustrated in Figure 3 and Figure 4. The first is a *global-context* approach whereby the KBM is learned with data from the entire test sequence. This approach follows the algorithm described in Section 2.1. The second is a variant referred to as a *local-context* approach whereby the KBM is learned from a shorter context window centred on the hypothesised speaker change point. Unlike the global-context approach, the local-context approach uses all the Gaussians contained in the defined context (no selection process is performed). This approach to KBM learning enables the flexible use of acoustic context information.

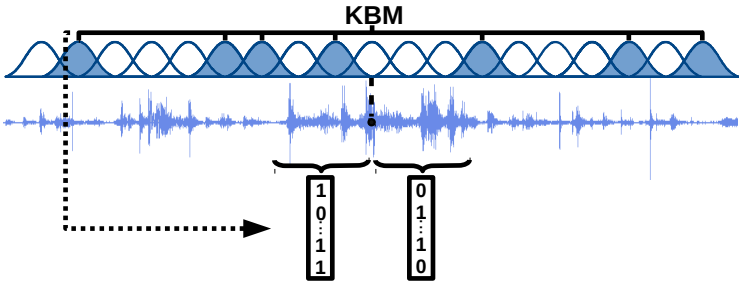


**Fig. 3. Global-context KBM** obtained through the selection of Gaussians from a global pool

### 3.2 Speaker change detection

Speaker change detection (SCD) is performed using data from two smaller and non-overlapping windows, one either side of hypothesized speaker change points. BKs are extracted for each window and are compared using the Jaccard distance, defined as:

$$D_J(\mathrm{v}_a, \mathrm{v}_b) = 1 - S_J(v_a, v_b) \tag{1}$$

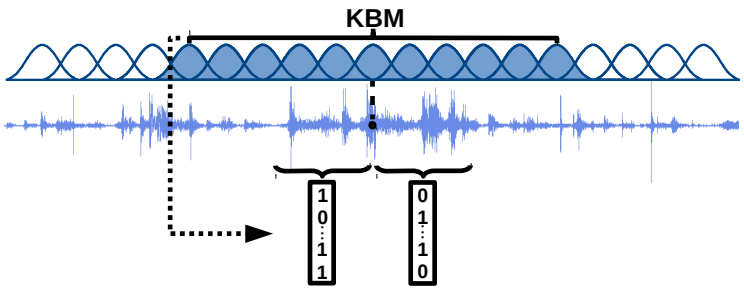


**Fig. 4. Local-context KBM** constructed using all Gaussians estimated from within a local context

where $S_J(\mathrm{v}_a, \mathrm{v}_b)$ is the Jaccard similarity between two binary vectors $v_a$ and $v_b$ defined as:

$$S_J(\mathrm{v}_a, \mathrm{v}_b) = \frac{\sum_{i=1}^{N}(v_a[i] \wedge v_b[i])}{\sum_{i=1}^{N}(v_a[i] \vee v_b[i])} \tag{2}$$

where $N$ is the vector dimension, $\wedge$ is the boolean AND operator and $\vee$ is the boolean OR operator.

This procedure is applied sequentially to obtain a curve of window distances at regular intervals. Local peaks in this curve represent speaker change candidates. Speaker change decisions are then obtained by thresholding the distance curve using an empirically optimised threshold.

## 4 Experimental setup

This section describes the database, the configuration of baseline and BK-based approaches to SCD and the evaluation metrics.

### 4.1 Database

In keeping with previous work on SCD, e.g. [17, 6], this work was performed with the ETAPE database [16] which contains audio recordings of a set of French TV and radio shows. The TV show development partition used for all work reported here consists of 9 audio files containing excerpts of debate, entertainment and broadcast TV shows broadcast on a number of different French TV channels. Together the recordings contain in the order of 5.5 hours of audio of which 3.9 hours contain the speech of 61 speakers in 2304 speech segments.

### 4.2 Baseline system

Acoustic features comprise 19 static Mel-frequency cepstral coefficients (MFCCs) which are extracted from pre-emphasised audio signals using an analysis window

of 25ms with a time shift of 10ms using a 20-channel Mel-scaled filterbank. No dynamic features are used.

The baseline SCD approach is a standard BIC segmentation algorithm [8]. It is applied with two windows of 1s duration either side of a hypothesised speaker change point. The resulting BIC distance curve is smoothed by replacing each point with the average estimated over 1s context. Local maxima are identified by enforcing a minimum distance of 0.5s between consecutive peaks. Within any 0.5s interval, only the highest peak is retained before speaker change points are selected by thresholding. This is a standard approach similar to those reported in [10, 7, 9].

### 4.3 Binary key system

Acoustic features are the same as for the baseline system. Candidate Gaussians for the KBM pool are learned from windows of 2s duration with a time shift of 1s. The number of components in the final KBM is chosen adaptively according to a percentage $\alpha$ of the number in the initial pool. Reported below are a set of experiments used to optimise $\alpha$. The number of top Gaussians $N_G$ used for BK extraction (step 5 in Figure 1) is set to 5 and the number of bits $M$ that are set to 1 is set to 20% of the number of KBM components $N$.

Two BKs are extracted every 0.1s with sliding windows of 1s duration positioned either side of the hypothesized change point. The distance between each pair of BKs is calculated using the Jaccard similarity (Section 3.2) and the distance curve is smoothed in the same way as for the baseline system. Speaker change points are again selected by thresholding.

### 4.4 Evaluation metrics

SCD performance is evaluated using the approach used in [6], namely through estimates of segment coverage and purity. According to the work in [6], coverage is defined as:

$$\text{coverage}(\mathcal{R}, \mathcal{H}) = \frac{\sum_{r \in \mathcal{R}} \max_{h \in \mathcal{H}} |r \cap h|}{\sum_{r \in \mathcal{R}} |r|} \tag{3}$$

where $|r|$ is the duration of segment $r$ within the set of reference segments $\mathcal{R}$, and where $r \cap h$ is the intersection of segments $r$ and segments $h$ within the set of hypothesis segments $\mathcal{H}$. Purity is analogously defined with $\mathcal{R}$ and $\mathcal{H}$ in Eq. 3 being interchanged.

An over-segmented hypothesis (too many speaker changes) implies a high segment purity at the expense of low coverage (hypothesised segments *cover* a low precentage of reference segments). In contrast, an under-segmented hypothesis (too few speaker changes) implies the opposite, namely high coverage, but low purity. Purity and coverage are hence a classical trade-off, with the optimal algorithm configuration depending on the subsequent task.

In order to concentrate on the assessment of SCD alone, ground-truth annotations are used for speech activity detection (SAD). It is noted that the use

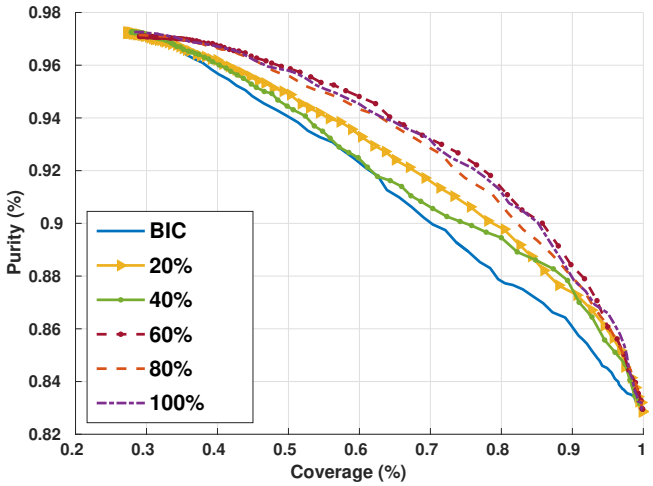


**Fig. 5.** Performance measured in segment purity and coverage using **global-context KBM**, obtained by varying the decision threshold $\theta$

of ground-truth SAD as a hypothesis with a single speaker delivers a ceiling coverage of 100% and a floor purity of 83%. These values can be taken as a performance reference.

## 5 Results

Figures 5 and 6 show plots of purity and coverage for global- and local-context KBMs respectively. Each profile shows the trade-off between the two metrics as the distance threshold $\theta$ is varied. Profiles are shown for KBMs whose size $\alpha$ is set to 20, 40, 60, 80 and 100% of the total number of original Gaussians. In both cases, the performance of the BIC baseline system is illustrated with a solid blue line.

The BK approach with global-context KBMs (Figure 5) gives universally better performance than the baseline, even if the trend is somewhat inconsistent. This behaviour is due to the Gaussian selection process which can result in a selection of Gaussians that are not representative of certain audio segments. KBMs of larger size have inherently better potential to cover the full acoustic space and hence better potential to produce more discriminant BKs. Larger KBMs then give better performance, e.g. for $\alpha$ greater than 40%. The optimal $\alpha$ is 60%. Greater value of $\alpha$ do not necessarily give better performance. This is because the acoustic space is already fully covered and the introduction of additional Gaussians is largely redundant. The BK approach with local-context KBMs (Figure 6) also outperforms the baseline. While the trend is consistent for lower values of coverage, across the range the optimal $\alpha$ varies between 60% and 100%.

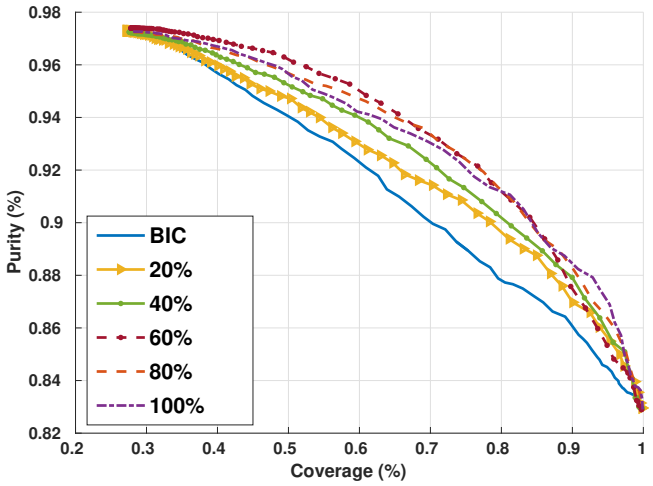


**Fig. 6.** Performance measured in segment purity and coverage using **local-context KBM**, obtained by varying the decision threshold $\theta$

Table 1 illustrates the variation in coverage against purity for BK-based SCD using global-context KBMs for $\alpha = 60\%$. The performance is compared to that obtained with the baseline system. The BK approach gives higher coverage at all operating points, especially for those with higher purity. Estimated using an area-under-the-curve (AUC) metric, the average relative improvement in coverage and purity across all operating points is 17.39% and 4.48%, for coverage and purity metrics respectively.

Table 2 illustrates the same analysis for BK-based SCD using local-context KBMs for $\alpha = 70$. Using the same AUC metric, the average relative improvement in coverage and purity is 18.71% and 4.51% respectively. These improvements are similar to results achieved by more advanced deep learning-based solutions such as that in [6].

It is of interest to compare the two proposed methods not only in terms of performance, but also in terms of efficiency and practical application. Even if the local-context approach slightly outperforms the global-context one, each approach can be better suited for different application modes. On one hand, in the case of offline processing (when the entire input stream is available in advance), the global-context approach is more efficient since the KBM is fix for all the process, hence allowing to compute frame-wise likelihoods only once and then reuse them for subsequent operations. However, in the local-context approach, the KBM changes over time (by using Gaussian components estimated on the contextual window around the current time point). This forces to recompute frame-wise likelihoods every time the window is shifted, therefore adding an extra computation cost. On the other hand, in online processing scenarios, the global-context approach cannot be used since the complete input stream is required in advance to train the KBM. However, the local-context approach is well suited

for online applications since it only utilises local information around the current time point. In the latter case, system latency is proportional to the amount of contextual data considered.

**Table 1.** Comparative performance measured in coverage for several fixed purity values using the **global-context KBM** approach

| **Purity (%)** | | 84 | 88 | 92 | 96 |
|---|---|---|---|---|---|
| **Coverage (%)** | **BIC** | 96.48 | 79.54 | 60.92 | 37.90 |
| | **BK** | 98.88 | 91.71 | 78.46 | 48.99 |

**Table 2.** Comparative performance measured in coverage for several fixed purity values using the **local-context KBM** approach

| **Purity (%)** | | 84 | 88 | 92 | 96 |
|---|---|---|---|---|---|
| **Coverage (%)** | **BIC** | 96.48 | 79.54 | 60.92 | 37.90 |
| | **BK** | 98.99 | 92.86 | 77.45 | 51.51 |

## 6 Conclusions

This paper introduces a binary key (BK) solution to speaker change detection (SCD). The algorithm uses traditional acoustic features and a configurable quantity of contextual information captured through a binary key background model (KBM). Speaker-discriminative BKs are then extracted from the comparison of acoustic features to the KBM. The binarisation of acoustic features resembles a form of quantisation which helps to reduce noise and hence improve the robustness of subsequent SCD. The latter is performed by thresholding the distance between BKs extracted from two adjacent windows either side of hypothesized speaker change points. While not requiring the use of external data, two variants of the novel BK SCD algorithm are shown to outperform a baseline approach based on the classical Bayesian information criterion. Results obtained using a standard dataset show average relative improvements which compare favourably to results reported recently for more computationally demanding deep learning solutions.

## Acknowledgements

This work was supported through funding from the Agence Nationale de la Recherche (French research funding agency) in the context of the ODESSA

project (ANR-15-CE39-0010). The authors acknowledge Hervé Bredin's help in the evaluation of speaker change detection.

## References

1. Anguera, X., Bonastre, J.F.: A novel speaker binary key derived from anchor models. In: Proc. INTERSPEECH. pp. 2118–2121 (2010)
2. Anguera, X., Bonastre, J.F.: Fast speaker diarization based on binary keys. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4428–4431. IEEE (2011)
3. Anguera, X., Movellan, E., Ferrarons, M.: Emotions recognition using binary fingerprints. Proc. IberSPEECH (2012)
4. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech, and Language Processing 14(5), 1505–1512 (2006)
5. Bonastre, J.F., Miró, X.A., Sierra, G.H., Bousquet, P.M.: Speaker Modeling Using Local Binary Decisions. In: Proc. INTERSPEECH. pp. 13–16 (2011)
6. Bredin, H.: TristouNet: Triplet Loss for Speaker Turn Embedding. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. na. IEEE (2017)
7. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the BIC. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 6, pp. VI–537. IEEE (2003)
8. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Proc. DARPA Broadcast News Transcription and Understanding Workshop. vol. 8, pp. 127–132 (1998)
9. Cheng, S.S., Wang, H.M., Fu, H.C.: BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Transactions on Audio, Speech, and Language Processing 18(1), 141–157 (2010)
10. Delacourt, P., Wellekens, C.J.: DISTBIC: A speaker-based segmentation for audio data indexing. Speech communication 32(1), 111–126 (2000)
11. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Improved binary key speaker diarization system. In: Proc. 23rd European Signal Processing Conference (EUSIPCO). pp. 2087–2091 (Aug 2015)
12. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Global speaker clustering towards optimal stopping criterion in binary key speaker diarization. In: Advances in Speech and Language Technologies for Iberian Languages: Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings. pp. 59–68 (2014)
13. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEE Transactions on Audio, Speech and Language Processing 23(12), 2286–2297 (2015)
14. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Novel clustering selection criterion for fast binary key speaker diarization. In: Proc. INTERSPEECH. pp. 3091–3095. Dresden, Germany (2015)
15. Delgado, H., Fredouille, C., Serrano, J.: Towards a complete binary key system for the speaker diarization task. In: Proc. INTERSPEECH. pp. 572–576 (2014)
16. Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., Galibert, O.: 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. p. na (2012)
17. Gupta, V.: Speaker change point detection using deep neural nets. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4420–4424. IEEE (2015)
18. Hrúz, M., Zajíc, Z.: Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. na. IEEE (2017)
19. Luque, J., Anguera, X.: On the modeling of natural vocal emotion expressions through binary key. In: Proc. 22nd European Signal Processing Conference (EUSIPCO). pp. 1562–1566 (2014)
20. Malegaonkar, A.S., Ariyaeeinia, A.M., Sivakumaran, P.: Efficient speaker change detection using adapted Gaussian mixture models. IEEE Transactions on Audio, Speech, and Language Processing 15(6), 1859–1869 (2007)
21. Neri, L.V., Pinheiro, H.N.B., Ing Ren, T., da C., C.G.D., Adami, A.G.: Speaker Segmentation using I-vector in Meetings Domain. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. na. IEEE (2017)
22. Patino, J., Delgado, H., Evans, N., Anguera, X.: EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation. Proc. IberSPEECH (2016)
23. Wang, R., Gu, M., Li, L., Xu, M., Zheng, T.F.: Speaker Segmentation using Deep Speaker Vectors for Fast Speaker Change Scenarios. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. na. IEEE (2017)
24. Wu, T.Y., Lu, L., Chen, K., Zhang, H.: Universal Background Models for Real-time Speaker Change Detection. In: MMM. pp. 135–149 (2003)
25. Zajíc, Z., Kunešová, M., Radová, V.: Investigation of Segmentation in i-Vector Based Speaker Diarization of Telephone Speech. In: International Conference on Speech and Computer. pp. 411–418. Springer (2016)