Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization

Héctor Delgado[1]*, Xavier Anguera*[2]*, Corinne Fredouille*[3]*, Javier Serrano*[1]

[1]CAIAC, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain
[2]Sinkronigo S.L., Barcelona, Spain
[3]CERI/LIA, University of Avignon, France

hecdelflo@gmail.com, xanguera@gmail.com
corinne.fredouille@univ-avignon.fr, javier.serrano@uab.cat

Abstract

Speaker diarization has become an important building block in many speech-related systems. Given the great increase of audiovisual media, fast systems are required in order to process large amounts of data in a reasonable time. In this regard, the recently proposed speaker diarization system based on binary key speaker modeling provides a very fast alternative to state-of-the-art systems at the cost of a slight decrease in performance. This decrease is mainly due to drawbacks in the final clustering selection algorithm, which is far from returning the optimum clustering the system is actually able to generate. At the same time, we have identified potential points of our system which can be further sped up. This paper aims to face these two issues by first lightening the processing at the main identified bottleneck, and second by proposing an alternative clustering selection technique capable of providing near-optimum clustering outputs. Experimental results on the REPERE test database validate the effectiveness of the proposed improvements, obtaining a relative performance gain of 20% and execution times of 0.037 xRT (being xRT the Real-Time factor).

Index Terms: Speaker diarization, binary key, within-class sum of squares, elbow criterion, cosine distance.

1. Introduction

Speaker diarization is the task of segmenting an audio stream into speaker-homogeneous segments (speaker clustering) and grouping them into speaker clusters according to the speaker identities (speaker clustering). Speaker diarization has a number of applications, commonly as a pre-processing tool for further speech technologies. For example, speaker diarization is widely used by speech-to-text engines in order to adapt acoustic models to the voices of the particular speakers speaking in the audio stream being processed [1]. Speaker recognition, which usually deals with audio from single speakers can benefit of the speaker separation provided by speaker diarization in a multi-speaker environment [2]. In the area of audiovisual content indexing, knowing who spoke when provides an added value to other automatically extracted metadata [3].

Current state-of-the-art speaker diarization systems provide very good performance. However, accurate speaker diarization requires the application of several costly algorithms, usually within an iterative scheme. Commonly, a combination of Gaussian Mixture Models for speaker modeling, Bayesian Information Criterion (BIC) for speaker segmentation and cluster merging, Viterbi decoding for data assignment, and others, are used to effectively perform speaker diarization, but at a cost of long processing times (xRT above 1, being xRT the Real-Time factor). Given the increasing volume of audiovisual content being generated, fast speaker diarization is required in order to process such an amount of data in a reasonable time.

In this regard, a fast speaker diarization based on the binary key speaker modeling was proposed in [4]. This system provides a very fast alternative, presenting real-time factors around 0.1xRT with a slight decrease of performance on the NIST-RT databases of meeting recordings. Later in [5], the approach was further investigated and applied to broadcast TV data, obtaining similar results. These works demonstrated the potential of the binary key speaker modeling for the speaker diarization task. However, they also reported problems in the stopping criterion, as the performance of the returned solutions are commonly far from the optimum clustering the system is able to generate. This fact motivated researching on the final clustering selection algorithm. In [6], an alternative global clustering method is applied towards optimal stopping criterion. However, this work achieved improvements only in a theoretical framework and is difficult to apply in practice.

Additionally, although our binary key system is already quite fast, we have identified the main bottleneck in the training of a UBM-like model called Binary Key Background Model (KBM), needed to train the binary keys (see Section 2). The training process involves an iterative selection of single Gaussian models based on the KL2 (Symmetric Kullback-Leibler) distance with the aim of selecting the most discriminant ones. We wondered if it would be possible to further speed up the process by lighten this key part by means of other distance measures between Gaussians.

In this paper we set two main objectives. First, we propose to accelerate the KBM training process by proposing a faster yet effective similarity measure between Gaussian models. Second, we further investigate towards a suitable clustering selection technique within the binary key speaker diarization framework, and propose a novel criterion based on the Within-Class Sum-of-Squares (WCSS). Experimental results show the effectiveness of our proposals in both speeding up the process and improving performance of the clustering selection stage.

The paper is structured as follows: Section 2 gives an overview of the binary key based baseline diarization system. Section 3 proposes a mechanism to speed up the KBM training while preserving performance. Section 4 describes the proposed clustering selection algorithm. Section 5 provides experimental results and discussion. Finally, Section 6 concludes and proposes future work.

2. Overview of the binary key speaker diarization system

A complete description of the binary key diarization system used in this work is given in [5]. In this paper only a brief overview is done. The system consists of two main blocks. First, the acoustic block transforms the input signal data into a series of binary vectors called Binary Keys (BK). Second, the binary block performs an Agglomerative Hierarchical Clustering (AHC) over the BKs.

The conversion of a set of acoustic features into a BK is carried out thanks to a UBM-like model called Binary Key Background Model (KBM), which is trained using the test data itself. Using a sliding window of a certain length and shift, single Gaussian models are trained on the test data. Window parameters are set in order to obtain an initial pool of a certain number of Gaussians. Then, the N most discriminant components are selected in an iterative process in which the remaining components in the pool are globally compared to the already selected ones by means of the KL2 (Symmetric Kullback-Leibler) distance (consult [4] for further details). In each iteration, the most dissimilar component is selected, until reaching the desired number of components.

Once the KBM is trained, any set or sequence of input feature vectors can be converted into a Binary Key. A BK $\mathrm{v}_f = \{v_f[1], ..., v_f[N]\}, v_f[i] = \{0, 1\}$ is a binary vector whose dimension N is the number of components in the KBM. Setting a position $v_f[i]$ to 1 (TRUE) indicates that the i-th Gaussian of the KBM coexists in the same area of the acoustic space as the majority of the acoustic data being modeled. The BK can be obtained in two steps. First, for each feature vector, the first N_G components providing the highest likelihood are selected, and their identifiers stored. Second, a vector of integer values called Cumulative Vector (CV) is calculated. The CV stores how many times each Gaussian in the KBM has been selected as a top-scoring Gaussian for the whole feature set being converted. Then, the final BK is obtained by setting to 1 the M top positions of the CV. This method can be applied to any set of features, either a sequence of features from a short speech segment, or a feature set corresponding to a whole speaker cluster.

The last step before switching to the binary process step is the clustering initialization. In this paper we use a simple uniform cluster initialization by splitting the input data into N_{init} equal-sized chunks. Although extremely simple, this approach has been extensively used with success.

The binary block implements an AHC clustering approach. However, all operations are done with binary data, which makes the process much faster than using classic GMM-based approaches (see Section 5). First, BKs for the initial clusters are calculated using the method explained above. Then, the input data, previously converted into a sequence of BKs, are reassigned to the current clusters, by comparing input BKs to all current cluster BKs by using some similarity measure. In [4], a simple similarity metric between binary vectors based on Boolean operators was proposed. In this work we opt for using the cosine similarity between CVs, as it has been proved to outperform the use of BKs in [7].

Once data have been assigned, BKs/CVs are trained for the new clusters. Next, similarities between all cluster pairs are calculated, and the cluster pair with the highest score is merged, reducing the number of clusters by one. The iterative process is repeated until a single cluster containing all the input BKs/CVs is obtained. At the end of the process, the output clustering has to be selected from those generated among all iterations. This is done by calculating the student T-test T_s metric as explained in [4] to all clustering solutions. Then, the clustering which maximizes T_s is returned.

3. Speeding up the KBM training

After analyzing execution times of the different modules involved in our speaker diarization system, we have concluded that the main bottleneck of our approach is the KBM training stage, and more specifically, the process of Gaussian component selection. As described in section 2, the KL2 distance is used to measure similarities between Gaussian models. In each iteration, KL2 distance has to be computed for the last selected Gaussian and the remaining ones in the Gaussian pool. KL2 provides a measure of how different two probability distributions are. D_{KL2}, namely "Symmetric Kullback-Leibler Divergence", of distributions P and Q is defined by Equation 1 as

$$D_{KL2}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \quad (1)$$

where $D_{KL}(P||Q)$ is the Kullback-Leibler divergence of distributions P and Q. KL is a non-symmetric measure, thus the KL2 measure is used instead. D_{KL} for multivariate normal distributions is defined by equation 2 as

$$D_{KL}(P||Q) = \frac{1}{2}\left(\mathrm{Tr}(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^T\Sigma_Q^{-1}(\mu_Q - \mu_P) - k + \log\left(\frac{|\Sigma_P|}{|\Sigma_Q|}\right)\right) \quad (2)$$

where Σ_P, μ_P, Σ_Q, and μ_Q are the covariance matrices and mean vectors of distributions P and Q, respectively, and k is the dimension of the data.

As it can be seen in equation 2, computation of KL2 involves a series of matrix operations, including the computation of traces, inversions and determinants. KL2 has been commonly used as cluster similarity measure and for speaker segmentation in speaker diarization. However, we wondered if we could use a simpler and faster, yet useful, method for our purposes. As the aim of the iterative Gaussian selection process is to select the most discriminant and complementary between them, possibly calculating distances between the means (centroids of the distributions) of the Gaussians could be discriminant enough to select the most dissimilar components. Following this reasoning, we propose the use of the cosine distance between Gaussian mean vectors as similarity metric. The cosine distance $D_{cos}(a, b)$ is defined as $D_{cos}(a, b) = 1 - S_{cos}(a, b)$, where $S_{cos}(a, b)$ is the cosine similarity between two vectors a and b, defined by equation 3 as

$$S_{cos}(a, b) = \frac{a \cdot b}{\|a\|\|b\|} \quad (3)$$

The cosine similarity formulation is considerably simpler than KL2 one, and its computation is faster. Nevertheless, it is still pending to assess if the cosine similarity will be discriminant enough and suitable in our Gaussian selection algorithm. This is evaluated in section 5.

4. Final clustering selection

Baseline results showed that the weakest point of the entire binary key speaker diarization system is the best clustering selection technique. It has been reported that the T_s based algorithm is far from returning the optimum number of speakers in our

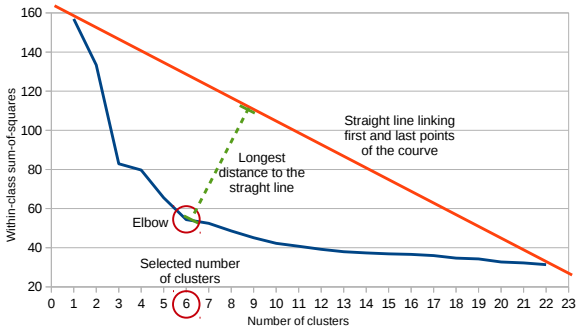

Figure 1: Example of the elbow criterion applied over the curve of within-class sum-of-squares per number of clusters. The point with longest distance to the straight line is considered the elbow.

system [4, 5]. It has also been shown that there exists a DER ceiling (i.e. DER of the optimum clustering selected manually), which has not been reached yet. This indicates that a better clustering selection will systematically result in an increase of performance.

In this paper, we propose a clustering selection technique based on the Within-Cluster Sum of Squares (WCSS). Given a clustering solution C_k composed of k clusters $c_1, c_2, ..., c_k$, each one containing multidimensional points (CVs or BKs) representing speech segments, the WCSS, $W(C_k)$, is defined as

$$W(C_k) = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2 \quad (4)$$

where μ_i is the mean of the points of cluster c_i (i.e. the centroid of cluster c_i). Actually, minimizing WCSS is the objective function used by the k-means clustering algorithm. However, WCSS can also be used as an indicator of how good a clustering solution is. Presumably, an accurate clustering solution originates clusters with small WCSS. Given a set of clustering solutions $C = (C_1, ..., C_{N_{init}})$, each one with an increasing number of clusters (from a single cluster to N_{init} clusters), WCSS can be calculated for all clustering solutions using Equation 4 and plotted as shown in Figure 3. When the number of clusters N is less than the optimum number of clusters, WCSS should be high. In the case of $N = 1$, WCSS is maximum, and when increasing the number of clusters, WCSS will follow an exponential decay. In some point the decay will become almost linear and WCSS will continue to fall smoothly. The first point which deviates from the exponential curve is considered as the elbow, and the associated number of clusters is selected as the optimum one. A simplified graphic approximation to find the elbow is to draw a straight line between the WCSS values of the first (with $N = 1$) and last ($N = N_{init}$) clustering solutions and calculate the distance between all the points in the curve and the straight line. The elbow is the point with the highest distance to the line. In the formulation of Equation 4, the euclidean distance of each cluster member (segment CVs) to its centroid (cluster CV) is used. However, we propose to use the cosine distance instead, which we found more suitable to compare CVs [7]. Therefore, the WCSS for a clustering solution C_k of k clusters, $W(C_k)$, can be reformulated as

$$W(C_k) = \sum_{i=1}^{k} \sum_{x \in c_i} (D_{cos}(x, \mu_i))^2 \quad (5)$$

5. Experiments and results

This section describes the experimental setup and results for two different experiments. In the first one, the new method of KBM estimation described in Section 3 is evaluated. Then, the second experiment assesses the clustering selection technique proposed in Section 4. Results of both experiments are compared to the ones obtained by the baseline system.

For speech activity detection, in this work we use ground-truth labels derived from the speaker diarization reference segments. This is done in order to evaluate the system without the impact of additional impurities introduced by false alarm speech. As for overlapping speech, we include such regions both diarization process and evaluation, although our system does not handle overlapping speech.

All tests are performed on the REPERE phase 1 test dataset of TV data [8]. This database was developed in the context of the REPERE Challenge [9]. It consists of a set of TV shows from several French TV channels.

5.1. Experimental Setup

Both experiments share a common configuration of system parameters, which is described next. In feature extraction, we extract standard 19-order MFCCs using a 25ms window every 10ms.

In the KBM training stage, an initial pool of 2000 Gaussians are trained on the test data itself using a sliding window of 2s with a shift which depends on the duration of the audio stream. Then, the final number of components is set to N components by following the Gaussian component selection algorithm explained in Section 2. In section 5.2, KL2 and cosine distance are evaluated and compared.

With regard to binary key estimate parameters, the top 5 Gaussian components are taken in a per frame basis, and the top 20% components at segment level.

For clustering initialization, we use a simple uniform initialization by splitting the input audio into N_{init} equal-sized chunks. Given the data being used, we set N_{init} to 25 initial clusters.

Finally, in the AHC stage, BKs/CVs keys are computed for each 1s segment, augmenting it 1s before and after, totaling 3s.

For performance evaluation, the output labels are compared with the reference ones to compute the DER. As said before, overlapping speech regions are included in both diarization processing and calculation of DER. In such regions with more than one speaker simultaneously, our system assigns only one speaker label.

5.2. KBM Training Experiments

Figure 2 shows the results obtained before ("BS", baseline) and after ("FAST KBM", improved) applying the improvements proposed for KBM training. In addition, for each system, two performance measures are given: DER of the best clustering produced by the system selected manually ("OptOut" in order to set a performance ceiling), and DER of the clustering returned by the final clustering selection technique ("SysOut"). With regard to execution time, xRT for each system is shown in the bottom part of Figure 2. Both DER and xRT are expressed as a function of the size of the KBM.

Execution time results confirm the hypothesis that using the cosine distance is faster than using the KL2 distance. In both cases, xRT increases linearly in function of the KBM size, but the slope obtained with the cosine distance is significantly

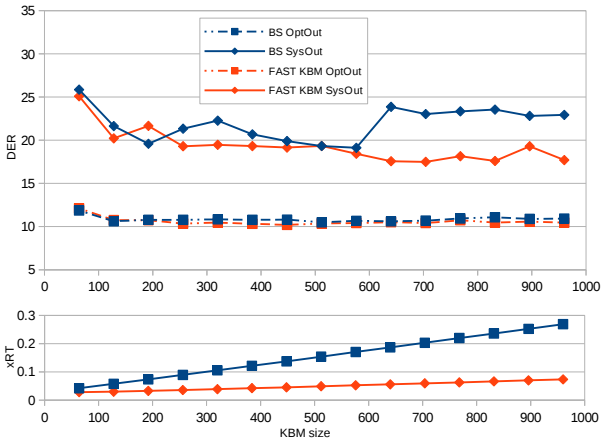

Figure 2: Performance and execution time measured in DER (top) and xRT (bottom), respectively, for the baseline and improved systems.

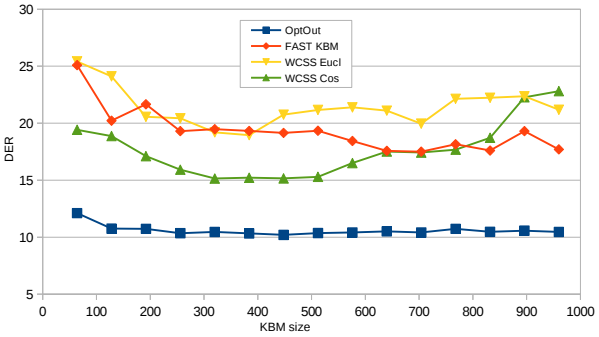

Figure 3: Performance evaluation of the newly proposed clustering selection based on the WCSS, measured in DER in function of the number of components in the KBM.

lower. The use of cosine distance between Gaussian components is not only faster than KL2, but also seems to be more accurate in the cases of the clustering selection output ("SysOut"). The best DER figure obtained by the baseline system is 19.12% using a KBM of 576 components, while the best performance by the improved system is 17.5% DER using a KBM size of 704. xRT values for those configurations are 0.170 and 0.059 for the baseline and the improved systems respectively. Although the best configuration of the improved system requires a bigger KBM than the baseline system, the sped up system is 2.87 times faster than the baseline, and 16.83 times faster than real time. Finally, as for performance ceiling (OptOut), both methods show very similar performance, reaching practically the same ceiling.

5.3. Final Clustering Selection Experiments

Figure 3 depicts the obtained results in the evaluation of the new proposed final clustering selection technique. In this experiment, we take as baseline the result "FAST KBM" obtained in the previous experiment. The improved system ("WCSS") replaces the T-test technique with the one proposed in Section 4. We have evaluated the approach using two different similarity measures between CVs in the calculation of the WCSS: the Euclidean distance ("WCSS Eucl", Equation 4), and the cosine distance ("WCSS Cos", Equation 5). First, it can be appreciated that the proposed criterion using the Euclidean distance does not outperform the baseline system. This probably occurs because the Euclidean distance could not be a suitable distance metric between Cumulative Vectors, in opposition with the T-test based criterion of the baseline system, which uses the similarity measure S between Binary Keys described in [4], which has been demonstrated to be a meaningful similarity metric between BKs [4, 5]. Second, the use of the cosine distance in the calculation of WCSS results beneficial, and the baseline system is finally outperformed. Finally, it can be seen that the new approach performs better with small sizes of the KBM, contrarily to the trend of the original selection algorithm, which performs better with bigger KBMs. This fact is also favorable in terms of execution time, as the more accurate modeling allows the use of smaller KBMs, with the subsequent improvement in system speed. The best performing configuration is the use of WCSS estimated with the cosine distance, and a KBM size of 320, providing a 15.15% DER. The relative improvement achieved is of 13.42% against the best performing configuration of the baseline (KBM of 704, resulting a 17.5% DER), and presents a real time factor of 0.037 xRT (25.6 times faster than real time).

6. Conclusions

This paper proposed improvements to our binary key speaker diarization system in two main aspects. First, KL2 distance was replaced with the cosine distance for measuring distances between Gaussian models within the iterative process of Gaussian selection of the KBM training algorithm. This aims to lighten the process and to achieve gains in execution time while preserving performance. Second, a novel clustering selection technique based on the calculation of the within-class sum-of-squares and elbow criterion was proposed in order to replace the current faulty clustering selection. Both improvements were validated experimentally, obtaining gains both in execution time and performance. In terms of performance, the resulting system achieves a relative improvement of 20%, with a final DER of 15.15%. In terms of execution time, the improved system runs 25.6 times faster than real time, obtaining 0.037 xRT computation time, and being 4.5 times faster than the baseline. Therefore, we have consolidated a very fast yet accurate speaker diarization system useful to process large audio collections. Obtained performance is not far from the obtained by the complex participating systems [10, 11, 12] in the official REPERE evaluation [13] on the same dataset, but surely, at a much lower computational cost. Although the new clustering selection improves performance significantly, there is still room for improvement, as the performance ceiling has not been reached yet. For this reason, we think it is worth further researching in this line. Finally, the proposed system should be evaluated on a different dataset to check if performance is stable across different speech data.

7. Acknowledgments

This work is part of the project "Linguistic and sensorial accessibility: technologies for voiceover and audio description", funded by the Spanish Ministerio de Economía y Competitividad (FFI2012-31024). This work was partially done within the French Research program ANR Project PERCOL (ANR 2010-CORD-102). This article is supported by the Catalan Government Grant Agency Ref. 2014SGR027.

8. References

[1] A. Stolcke, G. Friedland, and D. Imseng, "Leveraging speaker diarization for meeting recognition from distant microphones," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4390–4393.

[2] D. A. Reynolds and P. Torres-carrasquillo, "The mit lincoln laboratory rt-04f diarization systems: Applications to broadcast audio and telephone conversations," in *in Proc. Fall 2004 Rich Transcription Workshop (RT-04), Palisades*, 2004.

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

[4] X. Anguera and J.-F. Bonastre, "Fast speaker diarization based on binary keys," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4428–4431.

[5] H. Delgado, C. Fredouille, and J. Serrano, "Towards a complete binary key system for the speaker diarization task," in *INTERSPEECH*, 2014.

[6] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Global speaker clustering towards optimal stopping criterion in binary key speaker diarization," in *Advances in Speech and Language Technologies for Iberian Languages*, ser. Lecture Notes in Artificial Intelligence. Springer Berlin Heidelberg, 2014.

[7] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Improved binary key speaker diarization system," Submitted.

[8] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus : a multimodal corpus for person recognition," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.

[9] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly, "A presentation of the REPERE challenge," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, June 2012, pp. 1–6.

[10] M. Rouvier, G. Dupuy, P. Gay, E. el Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 1477–1481.

[11] H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V. B. Le, A. Roy, C. Barras, S. Rosset, A. K. Sarkar, Q. Yang, H. Gao, A. Mignon, J. J. Verbeek, L. Besacier, G. Quénot, H. K. Ekenel, and R. Stiefelhagen, "Qcompere @ repere 2013." in *SLAM@INTERSPEECH*, ser. CEUR Workshop Proceedings, G. Gravier and F. Béchet, Eds., vol. 1012. CEUR-WS.org, 2013, pp. 49–54.

[12] D. Charlet, C. Barras, and J.-S. Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates." in *ICASSP*. IEEE, 2013, pp. 7707–7711.

[13] O. Galibert and J. Kahn, "The first official repere evaluation." in *SLAM@ INTERSPEECH*, 2013.