EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation

Jose Patino[1], Héctor Delgado[1], Nicholas Evans[1] and Xavier Anguera[2]

[1]EURECOM, Sophia-Antipolis, France
[2]ELSA Corp., Barcelona, Spain
{patino,delgado,evans}@eurecom.fr,xanguera@gmail.com

Abstract. This paper describes the speaker diarization system submitted by EURECOM for the Albayzin 2016 speaker diarization evaluation. This evaluation consists of segmenting broadcast audio documents according to different speakers and attributing those segments to the speaker who uttered them, without any prior information about the speaker identities nor their number. EURECOM system is based on the binary key speaker modelling, an efficient and compact speech and speaker representation. The proposed system does not require external training data: the test data itself is used for estimating the resources needed. The system delivered a diarization error rate (DER) of 11.93% on the development set.

Keywords: speaker diarization, constant Q transform, ICMC, binary key, binary key background model

1 Introduction

Speaker diarization tries to answer the question of ‘who spoke when’ in an audio stream containing multiple speakers through segmenting and clustering speaker-homogeneous speech fragments. It is considered as an enabling technology that plays a key role in other subsequent tasks related to speech processing, such as automatic speech recognition, speaker recognition, speaker identification or spoken document retrieval. It still constitutes a very challenging problem for the speech processing community that needs of improvement in order to successfully tackle the increasing demand for real-life applications.

The Albayzin evaluation series has consolidated as a framework for promoting research on a number of speech processing tasks, such as audio segmentation, speaker diarization, text-to-speech, language recognition and spoken term detection. The last speaker diarization evaluation took part in 2010. After a period of 6 years, the speaker diarization evaluation has gained attention again, being one of the tasks evaluated in the present campaign. The main novelty of the current evaluation is the provision of speech activity detection (SAD) labels, thus the main focus of the evaluation is on the speaker-related errors rather than errors coming from SAD systems.

EURECOM is participating in the evaluation with two submissions. The systems presented are based on the binary key speaker modelling technique, a very efficient and compact way of modelling speakers. System configurations were tuned on the provided labelled development set. Then, the test data was processed with the best configurations found.

The paper is structured as follows: Section 2 gives an overview of the Albayzin 2016 Speaker Diarization Evaluation. Section 3 describes the speaker diarization system based on binary keys. Section 4 describes the experimental setup and results. Section 5 concludes and proposes future work.

2 Speaker Diarization Evaluation

This section briefly describes the Albayzin 2016 Speaker Diarization Evaluation. For a more detailed description refer to [8].

Its aim is contributing to the research in speaker diarization, which consists in segmenting audio files in homogeneous speaker turns to link them together according to the speaker identity. To ease the task, some information is provided beforehand. Specifically, speech, music and noise are labelled. Combinations of these three classes may occur creating complex situations of overlap that need to be addressed.

2.1 Database description

Audio files from various origins constitute the different data subsets for training, development, and testing.

Firstly, the training set, obtained from the Catalan broadcast news database from 3/24 TV channel, which was already used for the 2014 Albayzin Audio Segmentation Evaluation [12, 11] is provided. It was recorded by the TALP Research Centre from the UPC in 2009 under the Tecnoparla project [9]. The database contains approximately 87 hours of recordings of which speech constitutes roughly a 92%. Music and noise mean, respectively, a 20% and a 40% of the time. A last type classified as others accounts for a 3%. Finally, overlap is present in two different ways. 40% of speech time is overlapped with noise meanwhile a 15% is overlapped with music.

Secondly, the development and test sets are composed of files donated by the Corporacion Aragonesa de Radio y Television (CARTV). A total of approximately twenty hours selected from the Aragon Radio database have been split into two groups. One of four hours has been delivered as development set, where as the test set is composed of the remaining sixteen hours. Regarding its content, this second dataset is composed of around 85% speech, 62% music and 30% noise, where overlap is distributed as follows: a 35% of the audio contains music along with speech, a 13% overlaps speech with noise, and a 22% is constituted of speech alone.

All the data are supplied in PCM format, mono-channel, little endian 16 bit-per-sample, 16 kHz sampling rate.

2.2 Diarization Scoring

In order to evaluate the systems, the diarization error rate (DER) will be measured as the percentage of speaker time that is not rightfully assigned to a certain speaker. This scoring method, which follows the criterion applied at the NIST RT Diarization Evaluations [1], will be applied over the entire content of the files, without excluding overlapping regions, where more than one speaker are present.

Given a test dataset Ω, each document is divided into contiguous segments at all speaker change points, for both the reference and the hypothesis. Then, the diarization error time $E(n)$ is computed for each segment n as

$$E(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)] \tag{1}$$

where $T(n)$ is the duration of segment n, $N_{ref}(N)$ is the number of speakers that are present in segment n, $N_{sys}(n)$ is the number of system speakers that are present in segment n and $N_{correct}(n)$ is the number of reference speakers in segment n which are correctly assigned by the diarization system. Then, DER is calculated as

$$DER = \frac{\sum_{n\in\Omega} E(n)}{\sum_{n\in\Omega}(T(n)N_{ref}(n))} \tag{2}$$

Different kind of mistakes in the assignation of the speakers are considered in the diarization error time:

- **Speaker Error Time:** Considered as the amount of time wrongfully assigned to a speaker.
- **Missed Speech Time:** The Missed Speech Time makes reference to the amount of time where speech is present but is not labelled by the diarization system in segments where the number of system speakers is greater than the number of speakers in the references.
- **False Alarm Time:** This error time accounts for segments where speech is assigned by the system to a certain speaker but does not appear in segments where the number of speakers is greater than the number of speakers in the references.

Finally, to include in the criterion the possible mistakes introduced by human imprecision at the annotation of the files or ambiguity regarding starting and ending points for speech segments, a forgiveness collar of 0.25s, before and after each reference boundary, is considered.

3 Speaker diarization system

The speaker diarization system used in this evaluation is based on the system described in [5, 4]. It employs the so-called binary key (BK) speaker modelling approach, which offers a compact representation of a speech segment or cluster in the form of a vector containing zeros and ones. This vector captures speaker-specific characteristics and enables classification tasks by just computing similarity measures between BKs. Furthermore, its computation is very efficient compared with other state-of-the-art methods. The transformation is done by using a UBM-like model called binary key background model (KBM), which acts as a generator. Once the binary representation is derived from the input acoustic features, the subsequent operations are performed in the binary domain, and calculations mainly involve bit-wise operations between pairs of binary keys.

An overview of the speaker diarization system is depicted in Figure 1. The complete process consists of two stages. The first one, "acoustic processing", aims to map the input acoustic data into a sequence of BKs, while the second one, "diarization", performs the speaker diarization itself on the obtained binary representation. The two stages are described in detail below.

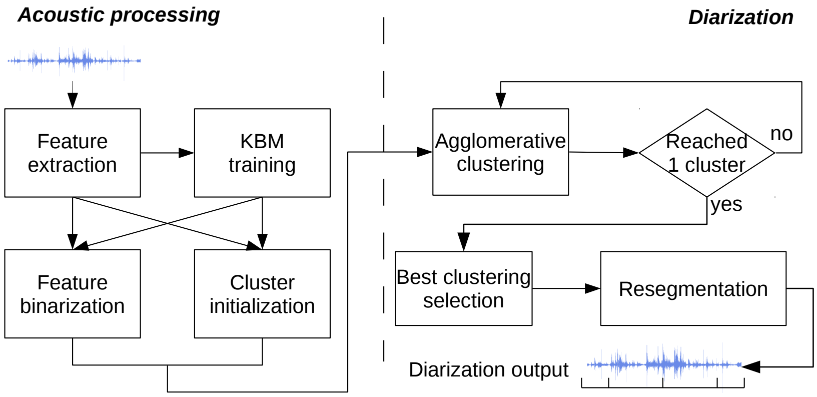

Fig. 1. Overview of the diarization system.

3.1 Acoustic processing

This stage performs the mapping of the input features into a sequence of binary keys. First, the input data is split into equal-sized segments using a sliding window with a certain shift (with some overlap between consecutive windows). From each segment, one binary key or cumulative vector is extracted. The resulting sequence will be the input data for the subsequent diarization process.

The binarisation of a sequence of acoustic features requires a KBM which is used as a generator model. Next, the KBM training and binary key extraction procedures are described.

KBM training. In order to estimate BKs, a UBM-like model called binary key background model (KBM) is required. This model is estimated on the test data stream itself, and therefore no external training data is required. The input feature stream is first windowed into frames of a given length with a given overlapping. Then, one single Gaussian model with diagonal covariance is trained on each data frame. This process results in an initial set or pool of single Gaussian components. At this point it is expected that many of those Gaussians are redundant. In order to select the most discriminant components and assure a good coverage for all the speakers, an iterative selection process is performed until the target number of components is reached. This process selects the most globally dissimilar Gaussian (from those not yet selected) to the ones already selected. The comparison of Gaussians is based on the cosine similarity between the Gaussian means (consult [4] for full details).

Binary key computation. Once the KBM has been trained, any set or sequence of acoustic feature vectors can be mapped into a binary key (BK). A BK $\mathrm{v}_f = \{v_f[1], ..., v_f[N]\}, v_f[i] = \{0, 1\}$ is a binary vector whose dimension n_{KBM} 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 acoustic data being modelled. The BK can be obtained in two steps, as it is shown in Figure 2. First, an initial binarisation at frame level is performed. Given one input feature vector, one binary vector is first initialized to zero. Then, the positions corresponding to the N_G top-scoring Gaussians (i.e. the Gaussians which provided the N_G highest likelihoods) are set to one. This binarisation is performed for each frame, resulting in a binary matrix with size n_f by n_{KBM}, being n_f the number of frames in the input sequence and being n_{KBM} the size of the KBM. Note that this vector can be efficiently computed by the partial sorting of the likelihoods for the current frame given by each Gaussian component of the KBM, and selecting their associated indices.

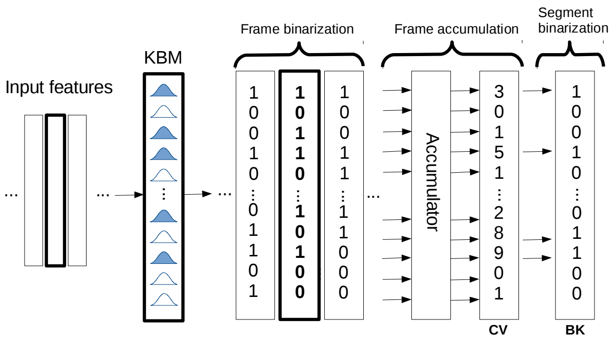

Fig. 2. Procedure for binary key extraction.

Second, the count of how many times each Gaussian has been selected as a top-scoring Gaussian along the input sequence of features is calculated, obtaining a cumulative vector (CV). The process is easily and efficiently implemented by summing the rows of the binary matrix obtained at frame level. Finally, the CV is processed to derive the BK by finding the top M Gaussians (i.e. the ones that were selected more frequently for the complete input feature sequence). Note that this procedure can be applied to an arbitrary set of features, either a sequence of features from a short audio segment, or a feature set corresponding to a complete speaker cluster. For classification tasks both the BK and the intermediate cumulative vector (CV) representation have been shown to be effective, depending on the application [7, 3]. In this work, the CV representation is adopted for the diarization of broadcast audio documents.

3.2 Diarization

The diarization process aims at clustering the sequence of CV into speaker clusters. The complete process is illustrated in Figure 3. First, an arbitrary number

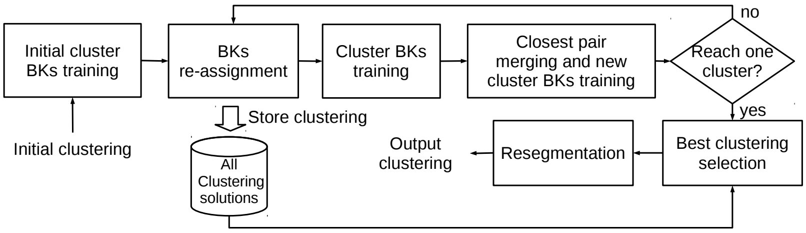

Fig. 3. Diarization process

of clusters N_{init} is initialized by just splitting the input sequence into N_{init} equal-sized segments and by estimating their corresponding cluster CVs. Then, a bottom-up, agglomerative hierarchical clustering (AHC) is performed. The input CVs are compared to the clusters' CVs by means of the cosine similarity. The cosine similarity returns a real number between 0 and 1, given that the CVs contain only positive values. Similarity values close to 0 indicate high dissimilarity, while values close to 1 indicate high similarity. Once the input CVs have been assigned to the clusters, the obtained clustering solution is stored. Then, cluster CVs are re-estimated with the new data partitions and compared in a pair-wise manner to find the most similar pair. Those clusters are then merged, forming a new cluster and therefore reducing the current number of clusters by one. The CV for the new cluster is estimated.

This process is repeated until one single cluster is obtained. At the end, a set of N_{init} solutions, each one with a decreasing number of clusters ranging from N_{init} to 1, is obtained. From the collection of solutions, the optimal one according to a certain criterion has to be selected. To that end, a criterion based on the trend in the within-cluster sum of squares (WCSS) among all the clustering solutions is employed. The main idea is to find a trade-off between the WCSS and the number of clusters. The solution is selected using the elbow method as described in [4].

Given that the system uses segments of fixed length to represent the input data in terms of CVs, the speaker segment boundaries may not be as precise as desired. In order to refine the segmentation, a final re-segmentation is performed on the solution returned by the clustering selection module. This re-segmentation relies on Gaussian mixture models (GMM) to model the clusters, and on maximum likelihood at acoustic feature level. A sliding window which moves through the feature sequence at a rate of 1 frame is evaluated against all cluster GMMs, and assigned to the one providing the maximum likelihood.

4 Experiments and results

Given an input audio file, the system must provide a set of segments including temporal information (beginning and duration) and speaker labels. In the case of overlapping speech, where more than one speaker speak simultaneously, the system should be capable of differentiating between the speakers and to annotate them properly. EURECOM's system does not include a dedicated module for overlapping speech detection. Therefore, it is expected that the system's miss speech error is close to the speaker time in overlapping regions.

The Albayzin 2016 Speaker Diarization Evaluation proposes two different training conditions. The open-set condition allows for external training data to be used, as long as it is publicly accessible. In a more restrictive manner, the closed-set condition limits the training data to that originally delivered by the organisers. EURECOM is participating in the later one, where closed-set constraints apply. An interesting characteristic of the proposed system is that it does not require any external training data, but it uses the test data itself for training the resources required. Despite the existence of a training set, it has not been used at all. The results on the development data reported on this paper, as well as the results submitted on the test set, depend uniquely on their own content.

One compulsory primary system and up to two contrastive systems can be submitted by each site. EURECOM is contributing with two submissions. They correspond to the same system, but employing different operation points.

In the following, the experimental setup is described. Later experimental results on the development set are reported. The final configurations for the official submissions are selected based upon the obtained results. Finally, execution time figures obtained when processing the evaluation set are provided.

4.1 Experimental setup

For feature extraction, the recently proposed infinite impulse response - constant Q, Mel-frequency cepstral coefficients (ICMC) [6] are used. These features employ the infinite impulse response - constant Q transform (IIR-CQT) time-frequency analysis tool [2]. IIR-CQT provides a multi-resolution spectrogram by IIR filtering of the fast Fourier transform (FFT), providing greater frequency resolution at low frequencies, and greater time resolution at high frequencies. 19 static cepstral coefficients are extracted from the pre-emphasised audio signal using a 25ms analysis frame, a shift of 10ms, a Hamming window and a 20-channel Mel-scaled filterbank and liftering [10].

As for KBM training, a 2s window with a rate of 0.5s is used to train the initial Gaussian pool. In order to avoid a small number of initial components for shorter audio files, a minimum amount of 1024 is forced (by decreasing the window shift conveniently). As regards the final size, and unlike in prior work where a unique KBM size was fix for a complete database [4], here the final size of the KBM is selected as a percentage of the initial pool size. In this way, the model size is chosen adaptively with regard to the audio file duration. The relative KBM size is swept across different percentages that go from the 5% to the 100% of the initial Gaussians sampled from the audio, in order to find the best configurations.

In the computation of CVs from the input data, segments of 1s augmented 1s after and before (totalling 3s), are considered. The number of top Gaussians per frame N_G is set to 5.

As for clustering initialisation, 25 initial clusters are derived from data chunks of equal size. This number is related to the maximum number of speakers found in the audio files of the development set (16 speakers at most).

4.2 Results

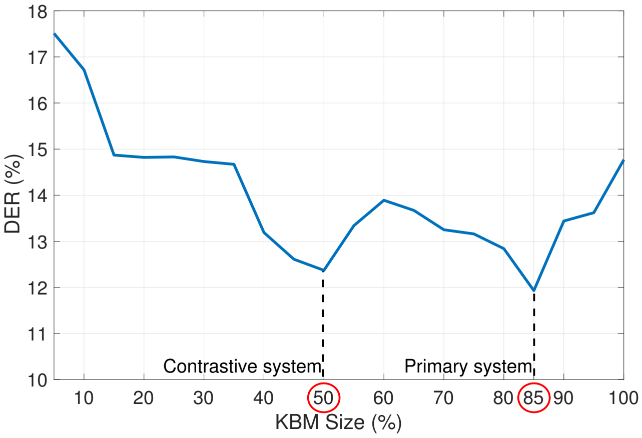

Fig. 4. DER trend for different KBM sizes on the development set.

Figure 4 shows system performance in terms of DER with regard to the relative KBM size. It is observed that there are two regions of interest where DER reaches minimum values. Those KBM sizes comprise the intervals 40-60% and 80-90%. The optimal DER of 11.93% is reached for a size of 85%. Given these results, operation points for the primary and contrastive submissions are selected at 85% and 50%, respectively.

Table 1 shows system performance for the chosen operation points on the development set. DER is broken-down into false alarm (FA), miss (MS), and speaker error(SE). Given that the ground-truth SAD labels are provided by the organisers, FA should be 0% and MS should be equal to the overlapping speech not detected by the system. However, FA is slightly above 0%. This can be due to imprecisions in segment boundaries returned. SE of contrastive system is just 0.4% above the primary system.

Table 1. Results obtained on the development set with the primary and contrastive configurations. FA stands for False Alarm, MS for Missed Speech, SE for Speaker Error and DER for Diarization Error Rate.

	KBM Size (%)	FA	MS	SE	DER
Primary system	0.85	0.5	2.0	9.4	11.93
Contrastive system	0.50	0.5	2.0	9.8	12.37

4.3 Processing the test set

Once the primary and contrastive systems were chosen from the development set, the test data was processed. Table 2 shows execution time figures obtained when running the systems on an Intel Core i5-3470 CPU desktop at 3.20GHz with 4 cores and 16 GB RAM, under a Ubuntu 14.04 operating system. For feature extraction only one single thread was used, while for speaker diarization 4 threads were employed. Given the lower dimension of the data, the contrastive system is more efficient than the primary one, with real-time factors (xRT) of 0.035 and 0.046, respectively.

Table 2. CPU time (hh:mm:ss) and real time factor (xRT) of primary and contrastive systems on the official test data.

	Primary system		Contrastive system	
Task	Time	xRT	Time	xRT
Feature extraction	00:49:11	0.046	00:49:11	0.046
Speaker diarization	00:39:59	0.044	00:32:01	0.035
Overall	01:29:10	0.090	01:21:12	0.081

5 Conclusions

This paper reported the EURECOM submissions to the Albayzin 2016 speaker diarization evaluation. The system submitted is based on the binary key speaker modelling, a compact and efficient representation of speech segments and speaker clusters. This system does not require any external training data, so the supplied training materials were not used at all. It will be of interest to compare performance with other systems which do employ training data. The proposed system obtained a DER of 11.93% on the development set, and a real time factor of 0.046xRT.

Acknowledgements

This work was partially done within the French Research program ANR Project ODESSA (ANR-15-CE39-0010).

References

1. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf
2. Cancela, P., Rocamora, M., López, E.: An efficient multi-resolution spectral transform for music analysis. In: Proc. ISMIR. pp. 309–314 (2009)
3. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Improved binary key speaker diarization system. In: Proc. EUSIPCO. pp. 2087–2091 (2015)
4. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Novel clustering selection criterion for fast binary key speaker diarization. In: Proc. INTERSPEECH. Dresden, Germany (2015)
5. Delgado, H., Fredouille, C., Serrano, J.: Towards a complete binary key system for the speaker diarization task. In: Proc. INTERSPEECH. Singapore (2014)
6. Delgado, H., Todisco, M., Sahidullah, M., Sarkar, A.K., Evans, N., Kinnunen, T., Tan, Z.H.: Further optimisations of constant Q cepstral processing for integrated utterance verification and text-dependent speaker verification (2016)
7. Hernández-Sierra, G., Bonastre, J.F., Calvo de Lara, J.: Speaker recognition using a binary representation and specificities models. In: Proc. CIARP. pp. 732–739. Argentina (2012)
8. Ortega, A., Vinals, I., Miguel, A., Lleida, E.: The Albayzin 2016 speaker diarization evaluation. In: Proc. IberSPEECH (2016)
9. Schulz, H., Costa-Jussa, M.R., Fonollosa, J.A.: Tecnoparla-speech technologies for catalan and its application to speech-to-speech translation. Procesamiento del lenguaje Natural 41, 319–320 (2008)
10. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: The HTK book. Cambridge university engineering department 3, 175 (2002)
11. Zelenák, M., Schulz, H., Hernando, J.: Speaker diarization of broadcast news in albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing 2012(1), 1–9 (2012)
12. Zelenák, M., Schulz, H., Hernando Pericás, F.J.: Albayzin 2010 evaluation campaign: speaker diarization. In: Proc. VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. pp. 301–304 (2010)