IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

EML Submission to Albayzin 2018 Speaker Diarization Challenge

Omid Ghahabi, Volker Fischer

EML European Media Laboratory GmbH, Berliner Straße 45, 69120 Heidelberg, Germany
omid.ghahabi@eml.org; volker.fischer@eml.org

Abstract

Speaker diarization, who is speaking when, is one of the most challenging tasks in speaker recognition, as usually no prior information is available about the identity and the number of the speakers in an audio recording. The task will be more challenging when there is some noise or music on the background and the speakers are changed more frequently. This usually happens in broadcast news conversations. In this paper, we use the EML speaker diarization system as a participation to the recent Albayzin Evaluation challenge. The EML system uses a real-time robust algorithm to make decision about the identity of the speakers approximately every 2 sec. The experimental results on about 16 hours of the developing data provided in the challenge show a reasonable accuracy of the system with a very low computational cost.

Index Terms: Speaker Diarization, Albayzin Evaluation, Online.

1. Introduction

Speaker diarization is the process of identifying who is speaking when during an audio recording. Compared to other speaker recognition tasks, speaker diarization is usually more difficult as there is no prior knowledge about the number and the identity of the speakers speaking in the audio recording. The task has received more attention by the speech community with increasing the number of broadcast, meeting, and call center recordings collected every year.

Albayzin evaluations are organized every two years and try to challenge the current unsolved problems in the speech processing area and are supported by the Spanish Thematic Network on Speech Technology (RTTH). One of the challenges this year has been speaker diarization of broadcast news recordings. The task itself is not new but it gets really challenging when the data is noisy, there is a background music when the speakers talk, and when speakers speak at the same time. Unlike the last Albayzin speaker diarization evaluation, the speech/nonspeech labels of the audio recordings are not available which makes the task more difficult.

In addition to the data of the previous evaluations, the new data is provided by the public Spanish National Television (RTVE) this year. The database comprises different programs broadcast by RTVE from 2015 to 2018. The programs cover a great variety of scenarios from studio to live broadcast, from read speech to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents [1].

Like in previous evaluations, two participating conditions are possible, close-set and open-set. In the close-set condition, only the data provided in the challenge is allowed to be used for the system development while for the open-set condition, the participating sites can use also other publicly available datasets in addition to the data provided in the challenge. We have participated only in the close-set condition in this paper.

We have used the EML Online speaker diarization system in this challenge. The system decides about the identity of the speakers at once every approximately 2 sec without any knowledge about the upcoming segments. The system uses the robust Voice Activity Detection (VAD) proposed in [2]. The algorithm used in the system is robust with a very low computational cost. The whole speaker diarization process including feature extraction is performed in an approximately $0.01 \times RT$.

The rest of the paper is organized as follows. Section 2 describes clearly the databases used to build the system. Section 3 describes briefly the algorithm and every part of the EML speaker diarization system. Section 4 explains the performance measurement metric and the software used for the evaluation. The experimental results are summarized in section 5. Section 6 concludes the paper.

2. Database

Three sets of data are provided in the challenge for training and development of speaker diarization (SD) systems. The first set is about 440 hours unlabeled broadcast news recordings. The second one is about 75 hours automatically labeled with another SD system. The last dataset is about 16 hours of human-revised labeled data. The final evaluation data set is also about 16 hours recordings from other channels. This data are collected from RTVE2018 [1], Aragon Radio, and 3/24 TV channel databases. The details are described as follows.

2.1. Unlabeled Data

The *train* partition of RTVE2018 [1] is a collection of TV shows drawn from diverse genres and broadcast by RTVE from 2015 to 2018. This partition is unlabeled and can be used for any evaluation task in Albayzin2018 [1]. The titles, duration and content of the shows included on the RTVE2018 database can be found in [1].

2.2. Automatically Labeled Data

This partition consists of the Aragon Radio, and 3/24 TV channel databases. The Aragon Radio database, which was donated by the Corporación Aragonesa de Radio y Televisión (CARTV), consists of around 20 hours of the Aragon Radio broadcast. About 35% of audio recordings contain music along with speech, 13% is noise along with speech and 22% is speech alone.

The 3/24 TV channel is a Catalan broadcast news database proposed for the Albayzin2010 Audio Segmentation Evaluation [3]. The owner of the multimedia content allows its use for technology research and development. The database consists of around 87 hours of recordings in which about 40% of the time speech can be found along with noise and 15% of the time speech along with music.

10.21437/IberSPEECH.2018-44

The Rich Transcription Time Marked (RTTM) files containing the segment information are generated automatically by another SD system and are provided in the challenge along with the audio recordings.

2.3. Human-Revised Labeled Data

The *dev2* partition of RTVE2018 [1] contains 12 audio recordings along with their human-revised RTTM files. This partition corresponds to two different debate shows, four programs (7:26 hours) of La noche en 24H, where a group of political analysts comments what has happened throughout the day, and eight programs (7:42 hours) of Millenium where a group of experts debates about a current issue. The audio recordings consist of speech, music, noise, and a combination of them.

3. EML Speaker Diarization System

We have used the EML Online speaker diarization system in which the audio recording is processed every 0.1 sec and the decision for the speaker ID is made approximately every 2 sec. In summary, every 0.1 sec, the Voice Activity Detection (VAD) algorithm decides if the current segment is speech or nonspeech. If it is nonspeech, it will be discarded otherwise the Baum-Welch statistics are computed and accumulated over speech segments until the predefined maximum speech duration is reached. Then the accumulated statistics are converted to the supervector which is further mean normalized by the UBM mean supervector. The resulting supervector is then converted to a lower dimensional speaker vector given a transformation matrix. The process of speaker vector extraction is shown in Fig. 1. Although the same feature vectors can be used for both VAD and supervectors, we have used two different ones referred to as VAD and SPK features in Fig. 1.

The speaker vector is compared with the current speaker models using cosine similarity. If the speaker vector does not belong to any speaker, based on a speaker-dependent threshold, a new model will be created. Before a new model creation, the speech segment is divided into two halves. For each half, a speaker vector is created and compared with another using cosine similarity. If the two halves are similar enough in terms of the speaker identity, the statistics are merged and the new model is created. Otherwise, each half is assigned to one of the current speaker models. In other words, a new speaker model is created only if two halves are similar enough. In the algorithm, every speaker has its own threshold which will be updated over time. All the new created speaker models are assigned a fixed starting threshold. Then this threshold is updated based on the average scores of the assigned speaker vectors over time. More details about every part of the algorithm is given as follows.

3.1. Feature Extraction and Universal background Models

Feature vectors are extracted every 10 msec with a 25 msec window. For VAD, they are 16 dimensional MFCCs along with their deltas, and for speaker vectors are 30 dimensional static MFCCs. Features are mean normalized with a 3 sec sliding window. UBM for both VAD and speaker vectors are GMMs with 64 Gaussian mixtures each. They are trained using the unlabeled dataset described in Sec. 2.1.

3.2. Voice Activity Detection

We have used the hybrid supervised/unsupervised VAD proposed in [2]. In summary, the VAD model is based on zero-order Baum-Welch statistics obtained from the UBM. Given the speech/nonspeech labels from the automatically labeled data (Sec. 2.2), a single 64 dimensional vector of zero-order statistics is obtained for each speech and nonspeech class. In the test phase, the zero-order statistics of a given audio segment is computed and compared with speech and nonspeech vectors based on the cosine similarity.

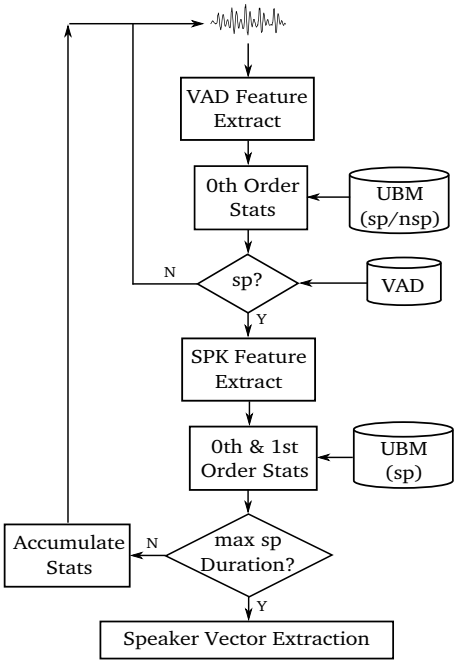

Figure 1: *Speaker vector extraction in the EML Online speaker diarization system.*

3.3. Speaker Vectors

Speaker vectors in this algorithm are obtained by the transformation of GMM supervectors. Supervectors are first transformed by a Within-Class Covariance Normalization (WCCN) matrix and then by a Linear Discriminant Analysis (LDA) matrix to lower dimensional vectors referred to as speaker vectors in this paper. The automatically labeled data (Sec. 2.2) is used to train WCCN and LDA. The whole data is divided into two non-overlapping datasets and chopped in 0.5, 1, 1.5, and 2 sec segments. For each segment one supervector is created. One dataset is used for training the WCCN and another for the LDA. We have used 300 dimensional speaker vectors in this challenge.

3.4. Scoring

The speaker vectors are compared using cosine similarity. However, the cosine score is not stable enough to make a robust decision. Therefore, we have created a set of background speaker vectors to normalize these scores effectively before decision. The background speaker vectors are obtained on the unlabeled dataset (sec. 2.1). All the speech segments are chopped into 2 sec segments and converted to 300 dimensional speaker vectors. Afterwards, the resulting speaker vectors are clustered using a two stage unsupervised clustering technique which was used to estimate the speaker labels of the background data for training the Probabilistic Linear Discriminant Analysis (PLDA) in [4].

Table 1: *The performance of the EML diarization system on the dev2 partition of RTVE2018 database.*

Response Time	**DER(%)**			**Total Processing Time**	**×RT**
	La noche 24H	Millenium	Total		
2 sec	32.65	12.24	22.12	00:12:37	0.014
4 sec	24.28	10.32	17.02	00:11:12	0.012

The first stage of the clustering algorithm is similar to the Mean Shift based algorithm proposed in [5] and used successfully in [6]. In the second stage, the closer clusters obtained in the first stage are combined. The second stage can be iterated for a few iterations or until no further merge is possible. In both stages, speaker vectors are joined based on the cosine similarity considering a threshold which is set to 0.350 and 0.300 for stages 1 and 2, respectively. After clustering, the centroids of the top 2048 clusters with higher number of members are considered as the final background speaker vectors.

Given the background speaker vectors, we perform a semi S-normalization on the scores before decision. The test speaker vector and the speaker models are both compared with background speaker vectors using cosine similarity. The cosine score between the test speaker vector and the speaker model is normalized once with the mean score of the top 10 closest background speaker vectors to the test speaker vector and another time with the mean score of the top 10 closest background speaker vectors to the speaker model. The final score is the average of these two normalized scores.

4. Performance Measurement

As in the NIST RT Diarization evaluations, the Diarization Error Rate (DER) will be used for the performance measurement in the challenge. The DER includes the time that is assigned to the wrong speaker, missed speech time and false alarm speech time.

The speaker error time is the amount of time that has been assigned to an incorrect speaker. This error can occur in segments where the number of system speakers is greater than the number of reference speakers, but also in segments where the number of system speakers is lower than the number of reference speakers whenever the number of system speakers and the number of reference speakers are greater than zero.

The missed speech time refers to the amount of time that speech is present but not labeled by the diarizaton system in segments where the number of system speakers is lower than the number of reference speakers.

The false alarm time is the amount of time that a speaker has been labeled by the diarization system but is not present in segments where the number of system speakers is greater than the number of reference speakers.

As defined in the challenge [7], consecutive segments of the same speaker with a silence of less that 2 sec come together and are considered as a single segment. A forgiveness collar of 0.25 sec, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a speaker begins or ends. Overlap regions where more than one speaker is present are also taken into account for the evaluation.

The tool used for evaluating the diarization systems is the one developed for the RT diarization evaluations by NIST md-eval-v22.pl, available in the web site of the evaluation: http://catedrartve.unizar.es/reto2018. The command line for the evaluation will be as follows,

```
md-eval-v22.pl -b -c 0.25 -r reference.rttm -s system.rttm  (1)
```

5. Experimental Results

We have used the human-revised labeled data (Sec. 2.3) for the evaluation of the diarization system. It contains 12 audio recordings from two different channels with a total duration of approximately 16 hours. The audio recordings include speech, music, silence, noise, cross talks (some times more than two speakers at a same time), and speech over music.

Two different participating conditions are proposed in this challenge, a closed-set condition in which only data provided within the Albayzin evaluation can be used for training and an open-set condition in which external data can also be used for training as long as they are publicly accessible to everyone. We have participated only in the closed-set condition.

The EML speaker diarization system is primarily designed for an Online application for which the robustness, computational cost, and the response time is important. In the primary submitted system, the decision about the identity of the speakers is made every approximately 2 sec without looking at the future in the audio recording. As the algorithm divides the speaker vectors into two halves before creating a new speaker model, the resolution for the speaker change point detection is about 1 sec. However, as the response time is not important in the challenge, we can increase it to a longer time but it would correspond to loosing fast speaker turns in the audio recording.

The development set (sec. 2.3) includes audio recordings from two Spanish programs, La noche en 24H and Millenium. The experimental results showed higher average DER on La noche en 24H recordings than on Millenium recordings. It could be due to longer duration of audio signals (2 hours each compared to 1 hour each for Millenium), faster speaker turns, more cross talks, more music on the background, or something else which needs more investigation.

Table 1 summarizes the average DER on each program, the total DER obtained on the entire *dev2* set of RTVE2018 dataset considering the duration of audio recordings, and the total computational time used for processing all the recordings from scratch including the feature extraction. The processing is made using a single core of an Intel(R) Xeon(R) CPU @2.10GHz.

6. Conclusions

We used the EML Online speaker diarization system as a participation in the recent Albayzin speaker diarization evaluation. We tried to take advantage of all the unlabeled and labeled data provided in the challenge in the close-set condition. The system showed a reasonable performance on the development data with a very low computational cost.

7. Acknowledgement

We would like to thank Wei Zhou for the efficient implementation of the Online algorithm.

8. References

[1] E. Lleida, A. Ortega, A. Miguel, V. Bazan, C. Perez, M. Zotano, and A. Prada, "RTVE2018 database description," 2018, [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf.

[2] O. Ghahabi, W. Zhou, and V. Fischer, "A robust voice activity detection for real-time automatic speech recognition," in *Proc. ESSV*, 2018.

[3] M. Zelenak, H. Schulz, and F. J. Hernando Pericas, "Albayzin 2010 evaluation campaign: speaker diarization," in *VI Jornadas en Tecnología del Habla*, 2010, pp. 301–304.

[4] O. Ghahabi and J. Hernando, "Deep learning backend for single and multisession i-vector speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 807–817, 2017.

[5] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 217–227, 2014.

[6] S. Novoselov, T. Pekhovsky, and K. Simonchik, "STC speaker recognition system for the nist i-vector challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 231–240.

[7] A. Ortega, I. Vinals, A. Miguel, E. Lleida, V. Bazan, C. Perez, M. Zotano, and A. Prada, "Albayzin evaluation: IberSpeech-RTVE 2018 speaker diarization challenge," 2018, [Online]. Available: http://catedrartve.unizar.es/reto2018/EvalPlan-SpeakerDiarization-v1.3.pdf.