Albayzin Evaluation: IberSPEECH-RTVE 2018 Speaker Diarization Challenge

Alfonso Ortega[1], Ignacio Viñals[1], Antonio Miguel[1], Eduardo Lleida[1], Virginia Bazán[2], Carmen Pérez[2], Manuel Gómez[2], and Alberto de Prada[2]

[1] Vivolab, Aragon Institute for Engineering Resarch (I3A)
University of Zaragoza, Spain
{ortega,ivinalsb,amiguel,lleida}@unizar.es
http://www.vivolab.es
[2] Corporación Radiotelevisión Española, Spain http:www.rtve.es

Abstract. The IberSPEECH-RTVE Speaker Diarization is a new challenge in the ALBAYZIN evaluation series. The evaluation is supported by the *Spanish Thematic Network on Speech Technology* (RTTH)[3] and *Cátedra RTVE Universidad de Zaragoza*[4] and is organized by ViVoLab Universidad de Zaragoza.The evaluation will be conducted as part of the Iberspeech 2018[5] conference to be held in Barcelona, Spain from 21 to 23 November 2018.
Speaker Diarization is a very important task for some speech technologies like Automatic Speech Recognition, Speaker Identification or Spoken Document Retrieval. This evaluation consists of segmenting broadcast audio documents according to different speakers and linking those segments which originate from the same speaker. No a priori knowledge is provided about the number or the identity of the speakers participating in the audio to be analyzed. The Diarization Error Rate will be used as scoring metric as defined in the RT evaluations organized by NIST. Two different conditions are proposed this year, a closed-set condition in which only data provided within this Albayzin evaluation can be used for training and an open-set condition in which external data can be used for training as long as they are publicly accessible to everyone (not necessarily free). Participants can submit systems in one or both conditions in an independent way.

1 Introduction

In some applications of speech technologies like Automatic Speech Recognition (ASR) systems for Broadcast shows, Speaker Identification (SPK ID) or Spoken Document Retrieval (SDR) in very large multimedia repositories, Speaker Diarization can be considered a very important task. Therefore, the development

[3] http://www.rthabla.es
[4] http://catedrartve.unizar.es
[5] http://iberspeech2018.talp.cat

of accurate Speaker Diarization Systems is essential to allow applications like ASR, SPK ID or SDR to perform adequately in real-world environments.

2 Database description and partitions

The Speaker Diarization evaluation consists of segmenting broadcast audio documents according to different speakers and linking those segments which originate from the same speaker. For this challenge, the evaluation database has been donated by Corporación Radiotelevisión Española (RTVE) and labeled thanks to the *Spanish Thematic Network on Speech Technology* (RTTH) and *Cátedra RTVE de la Universidad de Zaragoza*. Around sixteen hours of two different TV shows will be used for development and another sixteen hours from another two different shows will be used for testing. Also, the Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [1, 2] and the Corporación Aragonesa de Radio y Televisión (CARTV) database proposed for the 2016 Albayzin Speaker Diarization evaluation will be provided if needed for training purposes.

2.1 Training and Development data

RTVE2018 database. For training and development purposes, RTVE2018 database contains one training partition and a two development partitions. Around sixteen hours with diarization and reference speech segmentation will be distributed to registered participants in one of the development partitions and may be used for any purpose including system development or training[6]. The development data corresponds to two different debate shows, four programmes (7:26 hours) of *La noche en 24H*[7], where a group of political analysts comments what has happened throughout the day, and eight programmes (7:42 hours) of *Millenium*[8] where a group of experts debates about a current issue. The data will be distributed in AAC format, (LC mp4a), 44100 Hz, stereo, variable bitrate.[9]

Aragón Radio database. The database donated by the Corporación Aragonesa de Radio y Televisión (CARTV) consists of around twenty hours of the Aragon Radio broadcast. This data set contains around 85% of speech, 62% of music and 30% of noise in a way that 35% of the audio contains music along with speech, 13% is noise along with speech and 22% is speech alone. The data will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

[6] See RTVE2018 database description in `http://catedrartve.unizar.es/reto2018.html`

[7] http://www.rtve.es/alacarta/videos/la-noche-en-24-horas/

[8] http://www.rtve.es/alacarta/videos/millennium/

[9] We recommend ffmpeg to change to your audio format
https://www.ffmpeg.org/
ffmpeg -i file.aac -ar 16000 -ac 1 file.wav

3/24 TV channel database. The Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [1, 2] was recorded by the TALP Research Center from the UPC in 2009 under the Tecnoparla project [3] funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals (CCMA), owner of the multimedia content, allows its use for technology research and development. The database consists of around 87 hours of recordings in which speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called *others* was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music. The data will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

2.2 Evaluation data

RTVE database. The evaluation data will consist of around sixteen hours from a talk show and a debate show differents from the ones used for development. No a priori knowledge will be provided about the number or the identity of the speakers participating in the audio to be analyzed. The names of the shows and the broadcast dates will be provided at the time of data release. The data will be distributed in aac format, (LC mp4a), 44100 Hz, stereo, variable bitrate.[10]

3 Diarization Scoring

As in the NIST RT Diarization evaluations [4], to measure the performance of the proposed systems, the Diarization Error Rate (DER) will be computed as the fraction of speaker time that is not correctly attributed to that specific speaker. This score will be computed over the entire file to be processed; including regions where more than one speaker is present (overlap regions).

This score will be defined as the ratio of the overall diarization error time to the sum of the durations of the segments that are assigned to each class in the file.

Given the dataset to evaluate Ω, each document is divided into contiguous segments at all speaker change points found in both the reference and the hypothesis, and the diarization error time for each segment n is defined as

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

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 correctly assigned by the diarization system.

[10] We recommend ffmpeg to change to your audio format https://www.ffmpeg.org/

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

The diarization error time includes the time that is assigned to the wrong speaker, missed speech time and false alarm speech time:

- **Speaker Error 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.
- **Missed Speech Time**: 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.
- **False Alarm Time**: 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.

Consecutive segments of the same speaker with a silent of less that 2 seconds come together and are considered as a single segment. A forgiveness collar of 0.25 s, 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.

3.1 Segmentation Scoring Tool and Speaker Diarization System Output Files

The tool used for evaluating the segmentation system 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 format's definition for the submission of the Speaker Diarization results has been fixed according to the operation of the NIST's tool. Specifically the Rich Transcription Time Marked (RTTM) format will be used for speaker diarization system output and reference files. RTTM files are space-separated text files that contain meta-data 'Objects' that annotate elements of each recording and a detailed description of the format can be found in Appendix A of the 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan [4]. Thus, the required information for each segment will be:
SPEAKER File Channel Beg_Time Dur <NA> <NA> Speaker_Label <NA> <NA>
where:

- **SPEAKER**: A tag indicating that the segments contain information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker.
- **File**: It is the name of the considered file.
- **Channel**: It refers to the channel. Since we are dealing with mono recordings this value will always be 1.
- **Beginning_Time**: The beginning time of the segment, in seconds, measured from the start time of the file.
- **Duration**: It indicates the duration of the segment, in seconds.
- **Speaker_Label**: It refers to the label asigned to the speaker present in the considered segment.

The tag <NA> indicates that the rest of the fields are not used. The numerical representation must be in seconds and hundredth of a second. The decimal delimiter must be '.'.

The Albayzin 2018 Speaker Diarizaton evaluation will use the md-eval version 22 software. We will use the best match between hypothesis and reference labels. The command line will be:

md-eval-v22.pl -c 0.25 -b -r <SPKR-REFERENCE>.rttm -s <SYSTEM>.rttm

4 General Evaluation Conditions

The organizers encourage the participation of all researchers interested in speaker diarization. All teams willing to participate in this evaluation must send an e-mail to

- ortega@unizar.es
- lleida@unizar.es

Indicating the following Information:

- RESEARCH GROUP:
- INSTITUTION:
- CONTACT PERSON:
- E-MAIL:

with CC to Iberspeech 2018 Evaluation organizers at:

- albayzinevaluations@gmail.com

before September 24th, 2018.

4.1 Data License Agreement

The RTVE data is available to the evaluation participants only and subject to the terms of a license agreement with the RTVE. The license agreement can be downloaded from Cátedra RTVE-UZ web page:
http://catedrartve.unizar.es/reto2018.html

Participants must sign the agreement and send a scanned copy attached to the email. A copy signed by RTVE representative will be returned. Please read carefully the information provided on the Cátedra RTVE-UZ web page related with the use of the RTVE data after the evaluation campaign.

4.2 Evaluation Rules

Each participant team must submit at least a primary system in one condition, open-set or closed-set, but they can also submit up to two contrastive systems. Each and every submitted system must be applied to the whole test database. The ranking of the evaluation will be done according to results of the primary systems for the closed-set condition but the analysis of the results of the contrastive systems will be also processed and presented during the evaluation session at Iberspeech. All participant sites must agree to make their submissions (system output, system description, ...) available for experimental use by the rest of the participants and the organizing team.

The participant teams will notify and provide the total time required to run the set of tests for each submitted system (specifying the computational resources used). No manual intervention is allowed for each developed system to generate its output, thus, all developed systems must be fully automatic. Listening to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. The evaluated systems must use only audio signals. Any publicly available data can be used for training together with the data provided by the organization team to train the speaker diarization system only in the open-set condition. In case of using additional material, the participant will notify it and provide the references of this material. These databases must be publicly accessible although not necessarily free.

4.3 Result Submission Guidelines

The evaluation results must be presented in just one RTTM file per submitted system. The file output file must be identified by the following code:

EXP-ID::=<SITE>_<SYSID> where,

- <**SITE**>: Refers to a three letter acronym identifying the participant team (UPM, UPC, UVI, ...).
- <**SYSID**>: Is an alphanumeric string identifying the submitted system. For the primary system the SYSID string must begin with p-, c1- for contrastive system 1 and c2- for contrastive system 2.

Each participant team must send an e-mail with the corresponding RTTM result files to

- ortega@unizar.es
- lleida@unizar.es

4.4 System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fulfil the requirements given in the IberSpeech 2018 call for papers. You can use the templates provided for the Iberspeech conference (WORD or LaTeX). Please, include in your descriptions all the essential information to allow readers to understand the key aspects of your systems.

4.5 Schedule

- June 18, 2018: Registration opens. Release of the training and development data.
- September 24, 2018: Registration deadline. Release of the evaluation data.
- October 21, 2018: Deadline for submission of results and system descriptions.
- October 31, 2018: Results distributed to the participants.
- Iberspeech 2018 workshop: Official publication of the results.

5 Acknowledgments

The organizing team would like to thank to Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza by their efford for providing the data for the 2018 evaluation. Also, the Corporación Aragonesa de Radio y Televisión and Aragón Radio for providing the additional data for the evaluation. Thanks also to Martin Zelenak and Javier Hernando who organized the 2010 Albayzin Audio Segmentation Evaluation for their help, support and for providing the training material for this evaluation. And also to the organizing committee of Iberspeech 2018 for their help and support.

References

[1] Zelenak, M., Schulz, Hernando, J., Albayzin 2010 Evaluation Campaign: Speaker Diarization. VI Jorandas en Tecnologas del Habla, FALA 2010. Vigo, Noviembre 2010.

[2] Zelenak M., M., Schulz, Hernando, J., Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing. December 2012.

[3] Tecnoparla Project. Online: http://www.talp.upc.edu/tecnoparla, , accessed on June 2, 2016.

[4] The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, accessed on June 2, 2016.