Albayzin Evaluation: IberSPEECH-RTVE 2018 Multimodal Diarization Challenge

Eduardo Lleida[1], Alfonso Ortega[1], Antonio Miguel[1], Virginia Bazán[2], Carmen Pérez[2], Manuel Zotano[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. IberSPEECH-RTVE 2018 Multimodal Diarization evaluation is a new challenge in the ALBAYZIN evaluation series. The evaluation is supported by the *Spanish Thematic Network on Speech Technology* (RTTH) and *Cátedra RTVE Universidad de Zaragoza* and is organized by ViVoLab, Universidad de Zaragoza.The evaluation will be conducted as part of the Iberspeech 2018[3] conference to be held in Barcelona, Spain, from 21 to 23 November 2018.

The multimodal diarization evaluation consists of segmenting audiovisual documents according to different speakers and faces and linking those segments which originate from the same speaker and face.

1 Introduction

The multimodal diarization evaluation consists of segmenting broadcast audiovisual documents according to a closed set of different speakers and faces and linking those segments which originate from the same speaker and face. For this evaluation, a list of characters to recognize will be given. The rest of characters on the audiovisual document will be discarted for the evaluation purposes. System outputs must give for each segment who is speaking and who is/are in the image from the list of characters. For each character, a set of face pictures and short audiovisual document will be given.

The goal of this challenge is to start a new series of Albayzin evaluations based on multimodal information. This edition, we focus on face and speaker diarization. We want to evaluate the use of audiovisual information for speaker and face diarization. We encourage participants to use both speaker and face information jointly for diarization, although we accept systems that use visual and audio information separately.

[3] http://iberspeech2018.talp.cat

2 Database description

The RTVE2018 database donated by RTVE and labeled thanks to the *Spanish Thematic Network on Speech Technology* (RTTH) and *Cátedra RTVE de la Universidad de Zaragoza* will be used to evaluate multimodal diarization systems.

RTVE2018[4] database has been divided into 4 partitions, a *train* one, two development partitions *dev1*, *dev2* and finally a *test* partition.
Detailed information about the RTVE2018 database content can be found in the RTVE2018 database description report `http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf`. Only part of the *dev2* and *text* partitions will be used in this challenge.

Development data. For development, *dev2* partition contains a 2 hours show "La noche en 24H" labeled with speaker and face timestamps. Enrollment files for the main characters are also provided. Enrollment files consist on jpeg pictures and short mp4 videos with the character speaking. Addtionally, the *dev2* partition contains around 14 hours of speaker diarization timestamps. The *dev2* video files can be distributed to those participants who request it. We will appreciate to share face diarization timestamps if any participant create them.

Evaluation data. The evaluation data will contain a set of TV shows covering a variety of scenarios. We plan to have up to 8 hours of TV shows labeled with speaker and face timestamps. The detailed information about the *test* partition will be released along the evaluation data by September 24.

3 Performance Scoring

The multimodal diarization performance scoring will evaluate the accuracy of indexing a TV show in terms of the people speaking and present in the image. To measure the performance of the proposed systems, the Diarization Error Rate (DER) will be computed as the fraction of speaker or face time that is not correctly attributed to that specific character. This score will be computed over the entire file to be processed; including regions where more than one character 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 and face change points found in both the reference and the hypothesis, and the diarization error time for each segment n is defined as

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

[4] `http://catedrartve.unizar.es/reto2018.html`

where $T(n)$ is the duration of segment n, $N_{ref}(n)$ is the number of speakers or faces that are present in segment n, $N_{sys}(n)$ is the number of system speakers or faces that are present in segment n and $N_{Correct}(n)$ is the number of reference speakers or faces in segment n correctly assigned by the diarization system.

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

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

- **Speaker/Face Error Time**: The Speaker/Face Error Time is the amount of time that has been assigned to an incorrect speaker/face. This error can occur in segments where the number of system speakers/faces is greater than the number of reference speakers/faces, but also in segments where the number of system speakers/faces is lower than the number of reference speakers/faces whenever the number of system speakers/faces and the number of reference speakers/faces are greater than zero.
- **Missed Speech/Face Time**: The Missed Speech/face Time refers to the amount of time that speech/face is present but not labeled by the diarization system in segments where the number of system speakers/faces is lower than the number of reference speakers/faces.
- **False Alarm Time**: The False Alarm Time is the amount of time that a speaker/face has been labeled by the diarization system but is not present in segments where the number of system speakers/faces is greater than the number of reference speakers/faces.

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/face begins or ends.

The primary metric to rank systems will be the average of the face and speaker diarization errors

$$DER_{total} = 0.5 DER_{spk} + 0.5 DER_{face} \tag{3}$$

3.1 Segmentation Scoring Tool and Multimodal 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 *scoring* folder of the RTVE2018 database distribution.

The format's definition for the submission of the Multimodal 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 multimodal diarization system output and reference files. The RTTM files are space-separated text files that contains meta-data "Objects" that annotate elements of the recording. Each line represents the annotation of 1 instance of an object. Object types can be used or not depending on the particular evaluation. Table 1 shows the RTTM field names and values used in the RTVE2018 database. A more detailed description of the format can be found in Appendix C of the 2015 KeyWord Search Evaluation Plan[5]. For the sake of clarity an object named FACE has been defined to annotate the face appearances as SPEAKER is used for speakers turns annotation.

Table 1. RTTM files names used

Field 1	2	3	4	5	6	7	8	9	10
SPKR-INFO	file	1	<NA>	<NA>	<NA>	unknown	speaker_label face_label	<NA>	<NA>
SPEAKER	file	1	tbeg	tdur	<NA>	<NA>	speaker_label	<NA>	<NA>
FACE-INFO	file	1	<NA>	<NA>	<NA>	unknown	face_label	<NA>	<NA>
FACE	file	1	tbeg	tdur	<NA>	<NA>	face_label	<NA>	<NA>

Where:

- **SPEAKER/FACE**: A tag indicating that the segments contains information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker/face.
- **file**: It is the name of the considered file.
- **tbeg**: The beginning time of the segment, in seconds, measured from the start time of the file.
- **tdur**: It indicates the duration of the segment, in seconds.
- **Speaker/face_Label**: It refers to the label assigned to the speaker/face 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 Multimodal Diarizaton evaluation will use a modified version of md-eval version 22 software to accept the FACE object and the command line will be:

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

[5] https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf

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

- lleida@unizar.es
- ortega@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 licence 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 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 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 or watching to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. Any publicly available data can be used for training together with the data provided by the organization team. 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 Results Submission Guidelines

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

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

- <**SITE**>: Refers to the 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.
- <**MODAL**>: Is FACE for the output file of the Face Diarization System and SPKR for the output file of the Speaker Diarization System.

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

- lleida@unizar.es
- ortega@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 and release of the training 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 Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza for their efford providing the data for the 2018 evaluation. Thanks also to the organizing committee of Iberspeech 2018 for their help and support.