Albayzin Evaluation: IberSPEECH-RTVE 2022 Text and Speech Alignment Challenge: Alignment of re-spoken subtitles

Eduardo Lleida[1], Alfonso Ortega[1], Antonio Miguel[1], Virginia Bazán[2], Carmen Pérez[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

TaSAC-ST evaluation plan v1 (April 26, 2022)

Abstract. The IberSPEECH-RTVE 2022 Text and Speech Alignment challenge aims to evaluate the text to speech alignment systems on the actual problem of synchronising subtitles, generated by re-speaking, with the spoken audio at the subtitle level. Re-speaking is one of the most common approaches used for live subtitling. Re-speaking is a technique where a re-speaker listens to the original sound of a live programme and re-speaks it, including punctuation marks, to a speech recognition system. Two problems arise with this technique, the delay in displaying the subtitles and, most of the time, the re-speaker doesn't utter the same words that the original audio. 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 2022[3] conference to be held in Granada, Spain, from 14 to 16 November 2022.

1 Introduction

The IberSPEECH-RTVE 2022 Text and Speech Alignment Challenge aims to evaluate the text to speech alignment systems on the actual problem of synchronising re-speaking subtitles with the spoken audio. The task will assess the state of the art of offline alignment technology. The purpose is to provide the subtitles without delay for a new broadcast. In this task, participants will be supplied with the subtitles as they originally appeared on TV, including the start and end timestamps of each subtitle. Participants must provide an output with the exact same sequence of subtitles but with the new start and end timestamps for

[3] http://iberspeech2022.ugr.es

each subtitle. It should be noted that re-speaking subtitles often differs from the actual spoken words. If the speech is too fast the re-speaker tends to suppress words (deletions) or even paraphrasing, which introduce a new level of difficulty in the alignment process. The performance will be measure by computing the time differences between the aligned start and end timestamps given by the alignment systems and the reference timestamps derived from a careful manual alignment. A scoring script will be supplied to participants.

2 Challenge Description and Databases

The Text and Speech Alignment evaluation consists of automatically aligning the broadcast subtitles of different types of TV programs. For this evaluation, RTVE has licensed around 15 hours of three TV production jointly with the corresponding broadcast subtitles. The reference alignments have been manually supervised thanks to the Cátedra RTVE en la Universidad de Zaragoza.

2.1 Databases

2.1.1 RTVE2018DB

RTVE2018[4] database has a total of 569 hours and 22 minutes of audio. About 460 hours are provided with the subtitles and about 109 hours have been human-revised transcribed. Be aware that in most of the cases, subtitles could not contain a verbatim word transcription as most of them have been generated by a re-speaking procedure. The database has been divided into 4 partitions, a *train* one, two development partitions *dev1*, *dev2* and finally a *test* partition. Additionally, the database includes a set of text files extracted from all the subtitles broadcast by the RTVE 24H Channel during 2017.

2.1.2 RTVE2020DB

The RTVE2020 database is a collection of TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE) from 2018 to 2019. The database is composed of 55 hours of audio belonging to 15 different TV shows. The whole database has been human transcribed and it can be used as training/development partition for the Speech to Text Challenge.

2.1.3 RTVE2022DB

The RTVE2022 database is a collection of audio material belonging to 19 different genres. The recordings cover from old recordings of the RTVE archive, contest shows, social & cultural documentaries, unedited live interviews, newscast to present fiction series. For the Text to Speech Alignment Challenge, the RTVE2022DB contains a development partition with more than 2 hours of audio

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

and subtitles of 2 TV shows: Aquí la Tierra and Agroesfera. The test material will be made up of the same development TV shows and an additional new one with a total of 10 hours.

Detailed information about the RTVE databases content can be found in the RTVE database description reports. [5].

2.2 Training and Development data

For the IberSpeech-RTVE Text and Speech Alignment Challenge, participants are free to use whole RTVE2018DB and RTVE2020DB or any other data (speech and text) to train their acoustic and language models provided that these data are fully documented in the systems description paper. The **description of the training data must contains at least** the number of hours and origin of the speech data used to train the acoustic models and the size and origin of the text data used to train the language models. For public databases, the name of the database must be provided. For private databases, a brief description of the origin of the data must be provided.

As **development data**, we supply more than 2 hours of audio and subtitles of 2 TV shows, "Aquí la Tierra" and "Agroesfera".

2.2.1 Reference result.

As reference, we also provide the output of our baseline system. The reference system is based on a HMM speech recognized driven by a forced grammar at different levels. More details will be given in the results presentation.

2.3 Evaluation data

The evaluation data will contain new programs from "Aquí la Tierra" and "Agroesfera" and a new one from a total different genre. Around 10 hours of manually supervised aligned subtitles from the new RTVE2022DB will be used for evaluation. The detailed information about the evaluation data will be released by September 5th coinciding with the beginning of the evaluation task.

3 Performance Measurement

The alignment system output will be evaluated by a time-delay metric (TDM). All the participants will provide a file in stm format using the utf-8 charset per test file. The stm format describes the segment time marked files consisting of a concatenation of text segment records from a waveform file. Each record is separated by a newline and contains: the waveform's filename and channel identifier [A|B], the talkers ID, start and end timestamps (in seconds), optional subset

[5] `https://catedrartve.unizar.es/rtvedatabase.html`

label and the text for the segment. Here is an example of stm file:

20H 1 unknown 2079.102 2086.618 <,,> El premio se les concedió por sus descubrimientos sobre los mecanismos moleculares que controlan los ritmos cardiacos
20H 1 unknown 2086.642 2092.578 <,,> En la información que van a ver a continuación van a intentar explicar qué es exactamente eso .
20H 1 unknown 2093.900 2101.040 <,,> Los ritmos circadianos podrían traducirse popularmente como los mecanismos de nuestro reloj biológico interno

3.1 Average Program Time-Error Metric

The Average Program Time-Error Metric (APTEM) will be the primary metric for the Text and Speech alignment task. For each program in the test, a Program Time-Error Metric (PTEM) will be calculated and the final score will be computed by averaging the PTEM of each program in the test.

$$APTEM = \frac{1}{M} \sum_{m=1}^{M} PTEM(m) \tag{1}$$

where M is the number of programs in the test dataset.

3.1.1 Program Time-Error Metric

The Program Time-Error Metric (PTEM) is computed as follows. Given the probability distribution of the time difference between the reference and aligned start and end timestamps of each subtitle in a program, the PTEM is computed as the median value of the distribution.

Let's define the start-time error for the n-th subtitle, $TE_s(n)$, as

$$TE_s(n) = |Ts_{ref}(n) - Ts_{alig}(n)| \tag{2}$$

where, $Ts_{ref}(n)$ and $Ts_{alig}(n)$ are the start timestamps of the reference and the automatic aligned for the n-th subtitle, and the end-time error, $TE_e(n)$, as

$$TE_e(n) = |Te_{ref}(n) - Te_{alig}(n)| \tag{3}$$

where, $Te_{ref}(n)$ and $Te_{alig}(n)$ are the end timestamps of the reference and the automatic aligned for the n-th subtitle. Then, the time-error of the n-th subtitle, $TE(n)$ is computed as

$$TE(n) = TE_s(n) + TE_e(n) \tag{4}$$

The PTEM is defined as

$$PTEM = med([TE(1), TE(2), ..., TE(N)]) \tag{5}$$

where $med()$ is the median operator and N is the number of subtitles broadcast in the TV program.

The tool used to obtain the Program Time-Error Metric (PTEM) is “ptem.py´´ available at the scoring folder of the RTVE2022DB distribution. The command line is:

```
python ptem.py -r <SUBTITLES-REF>.stm -h <SYSTEM-HYP>.stm
```

The <SUBTITLES-REF>.stm and <SYSTEM-HYP>.stm must have the same correlative lines with the only diference of the start and end timestamps.

4 Evaluation Protocol

This challenge is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to the organizers for scoring.

4.1 Registration rules

The organizers encourage the participation of all researchers interested in text to speech alignment. All teams willing to participate in this evaluation must registered through the challenge web page
`http://catedrartve.unizar.es/albayzin2022.html`
before September 4th, 2022.
In case of any difficulty, you can send an e-mail to lleida@unizar.es

4.2 Data License Agreement

The RTVE data is available to the evaluation participants 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/rtvedatabase.html`

Participants must sign the agreement (digital signatures is valid) and send a 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.

4.3 Evaluation Rules

4.3.1 Submission procedure.

Each participant team must submit at least a primary system, but they can also submit up to three 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 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.

4.4 Results Submission Guidelines

The evaluation results must be presented in just one ZIP file per submitted system. The ZIP file must contain one TXT file per test audio file using utf-8 charset.
Each TXT file must be identified by the following code:
<FILENAME>_<SITE>_<SYSID>.txt
where,

- <**FILENAME**>: Refers to the filename of the test audio file without the extension (LM-20171215)
- <**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, c2- for contrastive system 2 and c3- for contrastive system 3.

The zip output file must be identified by the following code:
TASAC-ST_<SITE>_<SYSID>.zip

Each participant team must upload the zip files through the challenge web page
`http://catedrartve.unizar.es/albayzin2022.html`

In case of uploading problems send an e-mail with the corresponding ZIP result files to

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

4.5 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 2022 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.

A full conference paper can be submitted to the IberSpeech Conference as a regular paper for the Albayzin Evaluation special session. Please, take advise of the deadlines in the IberSpeech 2022 web page
`https://iberspeech2022.ugr.es/`

5 Schedule

- May 3rd, 2022: Registration opens and release of the training data.
- September 4th, 2022: Registration deadline.
- September 5th, 2022: Release of the evaluation data.
- October 16th, 2022: Deadline for submission of results and system descriptions.
- October 24th, 2022: Results distributed to the participants.
- October 30th, 2022: Paper submission deadline
- November 15th, 2022: IberSpeech 2022 special session in Granada.

6 Acknowledgments

The organizing team would like to thank Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza for their effort in providing the data for the 2022 evaluation. Thanks also to the organizing committee of Iberspeech 2022 for their help and support.