Albayzin Evaluation: IberSPEECH-RTVE 2020 Speech to Text Transcription Challenge

Eduardo Lleida[1], Alfonso Ortega[1], Antonio Miguel[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

S2T evaluation plan v1 (March 20, 2020)

Abstract. The IberSPEECH-RTVE 2020 Speech to Text Transcription aims to evaluate Automatic Speech Recognition (ASR) systems in realistic TV shows. 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 2020[3] conference to be held in Valladolid, Spain, from 18 to 20 November 2020.

1 Introduction

The IberSPEECH-RTVE 2020 Speech to Text Transcription Challenge aims to evaluate Automatic Speech Recognition (ASR) systems in realistic TV shows. The task will evaluate state of the art ASR technology to be used for applications as subtitling and automatic metadata generation for audiovisual content. Subtitling is the process by which we get a transcription of the audio portion of a program. Automatic metadata generation for audiovisual content is the process by which we analyze the content of the audiovisual document to archive, retrieve and filter audio-visual segments (for example, a special interview), objects (a special person) and events (a special goal in a football match)[1].

Tremendous progress has been observed during the last years in the performance of ASR systems. However they still entail errors, mainly due to challenging acoustic conditions, speaking rate, spontaneous speech, out-of-vocabulary words or language ambiguities. The resulting errors are of varying importance depending on the application in which the ASR system is being used. The most common measure of the ASR performance is the word error rate (WER). The WER is the edit distance between a reference word sequence and its automatic transcription. However, WER does not consider whether some words may be

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

more important to the meaning of the message. In fact, humans perceive different ASR errors as having different degrees of impact on a text. The ASR errors have different impact on both application, subtitling and automatic metadata generation. Usually, subtitling needs a closer verbatim transcription than automatic metadata generation as in the later the goal is to retrieve the relevant information present in the audiovisual document. These differences lead to different ways of measuring the performance of ASR systems. In this challenge, we will use word error rate (WER) as primary scoring measure but we will explore the use of other measures as Named Entity Error Rate (NEER). NEER is more suitable than WER for the evaluation of any application related with information retrieval from an audio document. A portion of the test material will be annotated in terms of Named Entities and the error rate will be computed using the annotated named entities as references.

2 Challenge Description and Databases

The Speech to Text transcription evaluation consists of automatically transcribe different types of TV shows. For this evaluation, RTVE has licensed around 640 hours of own TV production jointly with the corresponding subtitles. The shows cover a great variety of scenarios from scripted content to live broadcast, from read speech to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents including fiction series. Some of the contents have been labeled thanks to the Spanish Thematic Network on Speech Technology (RTTH) and 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 broadcasted by the RTVE 24H Channel during 2017.

Detailed information about the RTVE2018 database content can be found in the RTVE2018 database description report [5].

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

[4] `http://catedrartve.unizar.es/reto2018.html`
[5] http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

2019. The database is composed of 70 h and 18 min of audio belonging to 15 different TV shows. The whole database has been human transcribed and it will be used as test partition for the Speech to Text Challenge.

Detailed information about the RTVE2020 database content can be found in the RTVE2020 database description report [6].

2.2 Training and Development data

Participants can use the whole RTVE2018 database for training and development. In this new edition of the IberSpeech-RTVE Speech to Text Challenge, **we only consider the open training condition** where participants are free to use RTVE2018 training, development and test set 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 modesl. 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.

2.2.1 Reference result.

As reference of the performance, participants can use the previous IberSpeech-RTVE 2018 challenge results using as test the RTVE2018DB test partition. The results can be found in [2] [7].

2.3 Evaluation data

The evaluation data will contain a set of TV shows covering a variety of scenarios. Around 70 h of audio from the new RTVE2020DB will be used for evaluation. The detailed information about the evaluation data will be released by June 1st coinciding with the beginning of the evaluation task.

3 Performance Measurement

ASR system output will be evaluated with different metrics but a primary metric will be used for ranking ASR systems. All the participants will provide as ASR output for evaluation a free-form text with no page, paragraphs, sentence or speaker breaks with *.txt* extension using the utf-8 charset per test file. The text may include punctuation marks to be evaluated with an alternative metric. An example can be found in the *doc* folder of the RTVE2020 database.

[6] http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf
[7] https://www.mdpi.com/2076-3417/9/24/5412

3.1 Primary metric

Word Error Rate (WER) will be the primary metric for the Speech to Text Transcription task. The text will be normalized removing all the punctuation marks, numbers will be written with letters and text will be lowercased. The WER is defined as

$$WER = \frac{S + D + I}{N_r} \tag{1}$$

where N_r is the total words in the reference transcription, S is the number of substituted words in the automatic transcription, D is the number of words from the reference deleted in the automatic transcription and I is the number of words inserted in the automatic transcription not appearing in the reference. WER will be computed using the sclite tool included in the NIST Speech Recognition Scoring Toolkit (SCTK[8]). To use sclite tool it is necessary to translate the reference transcription files to any sclite reference format. Sclite accepts as reference files a variety of formats[9]. In this evaluation, we will use the *stm* format as reference. 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, begin and end times (in seconds), optional subset label and the text for the segment. The stm files are built from the transcription files (trn) using dummy segment time marks. Hypothesis files will be simply free-form text with no page, paragraphs, sentence or speaker breaks with *.txt* extension.
Here is an example of stm file:

20H 1 Presentador1 2079.102 2086.618 <,,> El premio se les concedió por sus descubrimientos sobre los mecanismos moleculares que controlan los ritmos cardiacos
20H 1 Presentador2 2086.642 2092.578 <,,> En la información que van a ver a continuación van a intentar explicar qué es exactamente eso .
20H 1 Voz_off8 2093.900 2101.040 <,,> Los ritmos circadianos podrían traducirse popularmente como los mecanismos de nuestro reloj biológico interno

3.2 Alternative metrics

In addition to the primary metric, other alternative metrics may be computed, but not taking into account for the challenge.

3.2.1 Punctuation marks evaluation (PWER)

The WER is computed with the punctuation marks given by the ASR system.

[8] https://www.nist.gov/itl/iad/mig/tools
[9] `http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`

3.2.2 Text Normalized Word Error Rate (TNWER)

Text normalization techniques as stopword removal and lemmatization are applied to the ASR output. In this sense, common errors as verbal conjugations, gender or number substitutions, articles, determiners, and quantifiers deletion/insertions will not have impact on the ASR performance. The same text normalization will be applied to both the reference and automatic transcriptions before proceeding to calculate WER. The freeling[10] lemmatizer will be used.

3.2.3 Named Entity Error Rate (NEER)

Part of the RTVE2020 database has been annotated in terms of Named Entities (NE). When dealing with information retrieval from audiovisual archives most of the search queries are related with named entities. For this reason, we will evaluate the performance of the ASR systems in terms of their capability to retrieve named entities from the audio signal. We will evaluate the named entity error rate by aligning the golden NE annotation with the output of the ASR. We will compute the NEER following the metric defined in the Message Understanding Conference (MUC)[11]. The following error will be considered [12]:

- Correct (COR) : a reference NE is in the hypothesis;
- Incorrect (INC) : the reference NE and the hypothesis don't match;
- Partial (PAR) : part of reference NE is in the hypothesis;
- Missing (MIS) : a reference NE is not in the hypothesis;
- Spurius (SPU) : the hypothesis doesn't exit in the reference NE;

The primary NEER metric is defined as:

$$NEER = \frac{INC + PAR/2 + MIS + SPU}{COR + PAR + INC + MIS + SPU} \tag{2}$$

The secondary NEE metric evaluates the error in terms of the number of correct predictions (COR), the number of actual predictions (ACT) and the number of possible predictions (POS), where

$$ACT = COR + PAR + INC + SPU \tag{3}$$

$$POS = COR + PAR + INC + MIS \tag{4}$$

Precision(P), Recall (R) and F-score are defined as:

$$P = \frac{COR}{ACT} \quad R = \frac{COR}{POS} \tag{5}$$

$$F - score = \frac{(\beta^2 + 1.0) \times P \times R}{(\beta^2 \times P) \times R} \tag{6}$$

where β is the relative importance given to recall over precision. $\beta = 1.0$, also known as F1-score, will be used to give the same importance to both recall and precision errors.

[10] `http://nlp.lsi.upc.edu/freeling/`
[11] https://en.wikipedia.org/wiki/Message_Understanding_Conference
[12] https://www.aclweb.org/anthology/M93-1007.pdf

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 speech to text transcription. All teams willing to participate in this evaluation must registered through the challenge web page
`http://catedrartve.unizar.es/albayzin2020.html`
before June 1st, 2020.
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:
<SITE>_<SYSID>.zip

Each participant team must upload the zip files through the challenge web page
`http://catedrartve.unizar.es/albayzin2020.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 2020 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, including test results and a post-analysis, can be submitted to the IberSpeech Conference as a regular paper. Please, take advise of the deadlines in the IberSpeech 2020 web page
`https://iberspeech2020.eca-simm.uva.es/`

5 Schedule

- March 23th, 2020: Registration opens and release of the training data.
- September 7th, 2020: Registration deadline. Release of the evaluation data.

- October 9th, 2020: Deadline for submission of results and system descriptions.
- October 30th, 2020: Results distributed to the participants.
- December 23th, 2020: Paper submission deadline
- March, 2021: Iberspeech 2020 conference.

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 2020 evaluation. Thanks also to the organizing committee of Iberspeech 2020 for their help and support.

References

[1] Kohler, J., Biatov, K., Larson, M., Eckes, C., Eickeler, S., "AGMA: Automatic Generation of Metadata for Audio-Visual Content in the Context of MPEG-7", Proceedings of Cast01, 2001.

[2] Lleida, E., Ortega A., Miguel A., Bazán-Gil, V., Pérez C., Gómez M., de Prada, A., Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media. Applied Sciences, Vol 9, Num 24, 2019, doi=10.3390/app9245412