Albayzin 2022 Text-to-Speech Alignment System Evaluation Subtask 2 - Evaluation Plan (v1, May 3, 2022)

Germán Bordel, Mikel Peñagarikano, Luis Javier Rodriguez-Fuentes, Amparo Varona

Grupo de Trabajo en Tecnologías Software (GTTS)
Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología, UPV/EHU
Barrio Sarriena s/n, 48940 Leioa

luisjavier.rodriguez@ehu.eus

Overview

Over the last years, with the widespread adoption of data-intensive deep learning approaches to ASR, the semi-supervised collection of training data for ASR has gained renewed interest. The Internet is plenty of resources pairing speech and text. Sometimes the paired text is an accurate transcription of the spoken content, sometimes it is only a loose and/or partial transcription, or even a translation to some other language. Therefore, a text-to-speech alignment system able to discriminate accurately paired speech and text segments becomes a very valuable tool.

With that goal in mind, the proposed task will deal with a long audio file, including sections in two different languages (Spanish and Basque), corresponding to a plenary session of the Basque Parliament. The paired text has been extracted from session's minutes and reflects only sections in Spanish. We aimed to focus on a single language and chose Spanish because most of the research groups aiming to participate in this evaluation would have ASR technology and resources for Spanish, but few would have them available for Basque. For syntactic correctness, sessions' minutes do not correspond exactly to audio content. Spontaneous speech events (such as filled pauses, false starts, repeated words, etc.) are ignored and some words are inserted, deleted or replaced just to make the text syntactically correct and easier to read. The audio parts in Basque are not expected to be paired with any text, but some words or word fragments (proper names, technical terms, etc.) may actually match (and be wrongly paired with) text in Spanish. Besides, the audio includes relatively long pauses between speaker turns and during voting times.

Text-to-speech alignment systems should compute, by automatic means, for each word in the text sequence (that is, in the same order as they appear), the start and end timestamps, a confidence score, and a hard Accept/Reject decision that would presumably be based on a score threshold. The alignment must be monotonous, that is, timestamps are always non-decreasing. The performance metric will only consider the accepted words, taking matching segments as positive and non-matching segments as negative, with the aim to reward those systems that are able to collect the highest amount of correctly transcribed audio with the lowest amount of incorrectly transcribed audio.

For development, the participants will receive an audio file (1 hour long) and the corresponding text to be paired with. A scoring script along with a ground-truth file will be also provided, so that the participants can score their alignments, obtaining a main performance score along with possibly other (secondary) performance measures and analyses. The ground-truth and the scoring script might be updated during the development phase, to fix potential issues, to account for other performance metrics or to extend the analyses. The development phase will span four months (May-August 2022) and should be employed to build and tune the text-to-speech alignment systems. Test data will be released on September 5th, 2022 and will consist of an audio file (1 hour long) along with the corresponding text to be paired with. Participants must submit their result files by October 16th, which must include alignments for both the development and test sets. Results for one primary system are mandatory. Besides, results for at most 4 contrastive systems can be submitted. Teams will be ranked according to the performance of their primary systems on the test set. Each participant will receive their performance results on October 18th, with no information about other participants' results.

The dataset

The audio data

Audio data is stored in 16 kHz 16-bit signed single-channel PCM WAV files. Audio recordings were originally made through the audio system (desktop microphones) of the Basque Parliament (BP) and are generally clear with high SNRs. Two different audio files (each approximately one hour long) will be provided for development and test. Both are extracts from the same plenary session, which features speech from several (not many) speakers, who may switch from Spanish to Basque (or viceversa) during their turns. Speaker turn changes and votings are both managed by the president of the BP and involve a certain amount of silent or slightly noisy regions, but speaker turn overlaps are very uncommon.

The paired text

The text to be aligned with the audios (hereafter, the paired text) has been extracted from the session's minutes. These minutes are based on the audio but ignore spontaneous speech events (such as filled pauses, false starts, repeated words, etc.) and include a sizable amount of editions to preserve syntactical correctness, which is frequently overlooked by speakers. As a consequence of this, the provided text does not match the audio, featuring word deletions, insertions and replacements. Sometimes, a word said in the audio is replaced in the minutes with a very similar variation of it (with different gender or number) and the most optimistic alignment will inescapably lead to an error, just because acoustics and spelling do not match. Both the paired text and the ground truth transcriptions have been normalized by removing punctuation marks, replacing accented vowels with non-accented vowels and converting all letters to lowercase. Uppercase letters have been kept only for acronyms (e.g. ADN, EH, UPyD, etc.) which could be either spelled (the most common case) or read as words. This should be taken into account when performing the alignment.

The paired text does not include the parts spoken in Basque, so there could be remarkable time leaps between one word (which may happen to be the end of a part spoken in Spanish) and the following word (which may happen to be the beginning of the next part spoken in Spanish, several minutes ahead in the audio signal). Again, this should be taken into account when performing the alignment.

The ground truth

The ground truth is based on manually generated rich text transcriptions, which include spontaneous speech events, such as filled pauses, false starts, cut words, etc. These transcriptions follow the acoustics even though the syntactical correctness is lost. The timestamps of sentences were manually added, so they are fully reliable. Word-level timestamps inside sentences were obtained automatically by forced alignment of each sentence transcription with the corresponding audio. To verify the accuracy of word-level timestamps, an informal test was carried out using several randomly chosen sentences, by manually adding word-level timestamps and comparing them with automatic segmentations. It was observed that differences between manual and automatic timestamps spanned from 0 to 20 milliseconds. Thus, the automatic segmentation was considered good enough for the purposes of this evaluation, provided that a reasonable collar time was applied.

For this evaluation, only the words appearing in the paired text are kept in the ground truth, the remaining elements of the rich text transcription being hidden. Note that we are interested only in how well the paired text is aligned with the audio. Taking this into account, if a word w in the paired text is aligned with an audio segment which is not included in the ground truth, we guarantee that neither the word w nor any other word in the paired text appears in that segment, so the time span of such segment is counted as error no matter the exact transcription of it. Also, to be fair with the participants, if a word w of the paired text (e.g. a proper name) appears in a part of the audio spoken in Basque, we include the corresponding segment in the ground truth, just to cover the case that the word w is aligned with that segment. Finally, to account for the uncertainty when defining the borders between words, a collar time can be established so that that a certain amount of time around the borders is not evaluated at all.

The task

The task consists of aligning each word of the text with a segment of the audio file so that the audio content corresponds to a pronunciation of the given word. Alignments must be monotonous, that is, the sequence of timestamps must be non-decreasing. Obviously, it is guaranteed that there is an optimal monotonous alignment between W and the audio signal X. Let $W = \{w_1, w_2, \dots, w_N\}$ be the sequence of N words to be aligned with an audio signal X, and $S = \{s_1, s_2, \dots, s_N\}$ the corresponding sequence of aligned segments in X. Then, if a word w_i is aligned to a segment $s_i = (t_1, t_2)$ and another word w_j is aligned to a segment $s_j = (t_3, t_4)$, with $i < j$, then the timestamps defining those segments must be $t_1 \leq t_2 \leq t_3 \leq t_4$. Non-monotonic alignments are not allowed and non-monotonic submissions will not be accepted.

The output of an alignment system must be a text file containing a line for each word in the paired text, each line including 5 columns (separated by any amount of spaces or tabs) with the following information:

- t_{beg}: a real number with the time when the segment starts.
- t_{end}: a real number with the time when the segments ends.
- word: the word paired with the audio segment.
- score: a real number reflecting the confidence on the alignment, the more positive the score, the higher the confidence; the more negative the score, the lower the confidence.
- decision: a 0/1 value, 0 meaning Reject and 1 meaning Accept. Remind that rejected words will not be evaluated.

The participants should develop one or more systems to automatically align the paired text with the audio, taking into account that some parts of the audio should not be aligned with any text and that the paired text does not reflects exactly the audio contents. It is not allowed to listen to the audio or use any kind of human intervention (e.g. crowdsourcing). Otherwise, any approach can be applied with no limit to the type or amount of resources that the participants can use to perform the task, as long as they describe the employed methods and resources with enough detail and, if possible, provide links to papers, data and/or software repositories that make it easier to reproduce their approach.

The performance metric

The ground truth is pre-processed before using it to compute the performance metric. First, the missing segments are added to the ground truth and assigned an Out-Of-Vocabulary label (‘#’). Then, the borders between segments are redefined by excluding from evaluation a collar time t_{collar} around them (in this evaluation, we are considering t_{collar} = 20 milliseconds): the starting time t_{beg} of each segment is added $t_{collar}/2$ while the ending time t_{end} of each segment is subtracted $t_{collar}/2$. These operations are suitably represented in Figure 1.

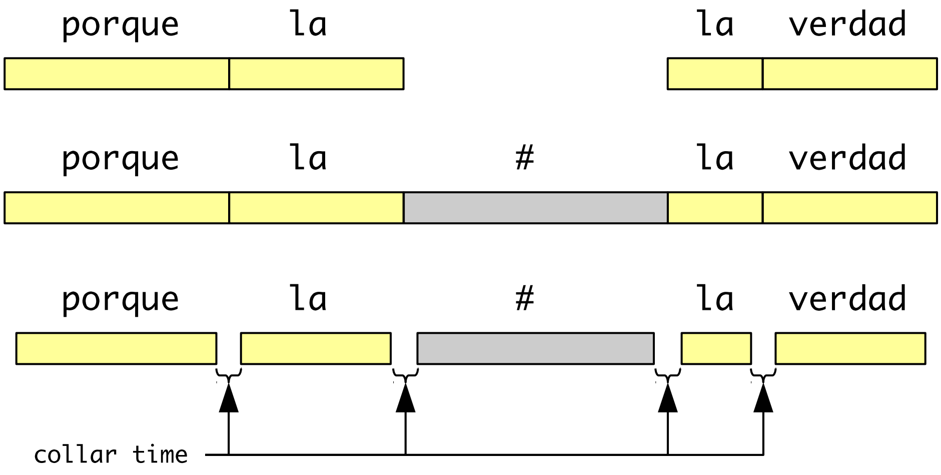

Figure 1. Ground truth pre-processing: first, the gaps in the ground truth are filled with an OOV label (‘#’); then, collar time is excluded from evaluation at the borders between segments.

Since the objective of the alignment is to recover as much correctly transcribed speech as possible to train acoustic models for the development of ASR systems, our performance metric should reflect this objective, but also the negative impact of wrongly aligned segments, because they could seriously compromise our semi-supervised training strategy. Thus, the performance metric will be just the difference between the correctly and the wrongly aligned times.

Let S = $\{s_1, s_2, \ldots, s_N\}$ be the alignment system output for the paired text W = $\{w_1, w_2, \ldots, w_N\}$. Only those segments accepted by the system will be evaluated, so we get the sequence of accepted segments S' = $\{s_1, s_2, \ldots, s_{N'}\}$ (with N' ≤ N). Each accepted segment is then aligned with the ground truth, which produces a sequence of sub-segments, each of them aligned either with a ground truth segment or with a collar time segment (see Figure 2). Sub-segments aligned with collar time are not evaluated and will not be considered hereafter. Let C = $\{c^{(1)}, c^{(2)}, \ldots, c^{(M)}\}$ be the sequence of sub-segments obtained after aligning the accepted segments with the ground truth, excluding collar-time. Each sub-segment is a 4-tuple:

$$c^{(i)} = \left(t_{beg}^{(i)}, t_{end}^{(i)}, w_a^{(i)}, w_g^{(i)}\right)$$

where $t_{beg}^{(i)}$ is the start time, $t_{end}^{(i)}$ is the ending time, $w_a^{(i)}$ is a word in the paired text and $w_g^{(i)}$ is a word in the ground truth. The performance metric is defined as follows:

$$score(C) = \sum_{i=1}^{M} \delta(w_a^{(i)}, w_g^{(i)}) \cdot \left(t_{end}^{(i)} - t_{beg}^{(i)}\right)$$

where:

$$\delta(w_a^{(i)}, w_g^{(i)}) = \begin{cases} 1 & w_a^{(i)} = w_g^{(i)} \\ -1 & w_a^{(i)} \neq w_g^{(i)} \end{cases}$$

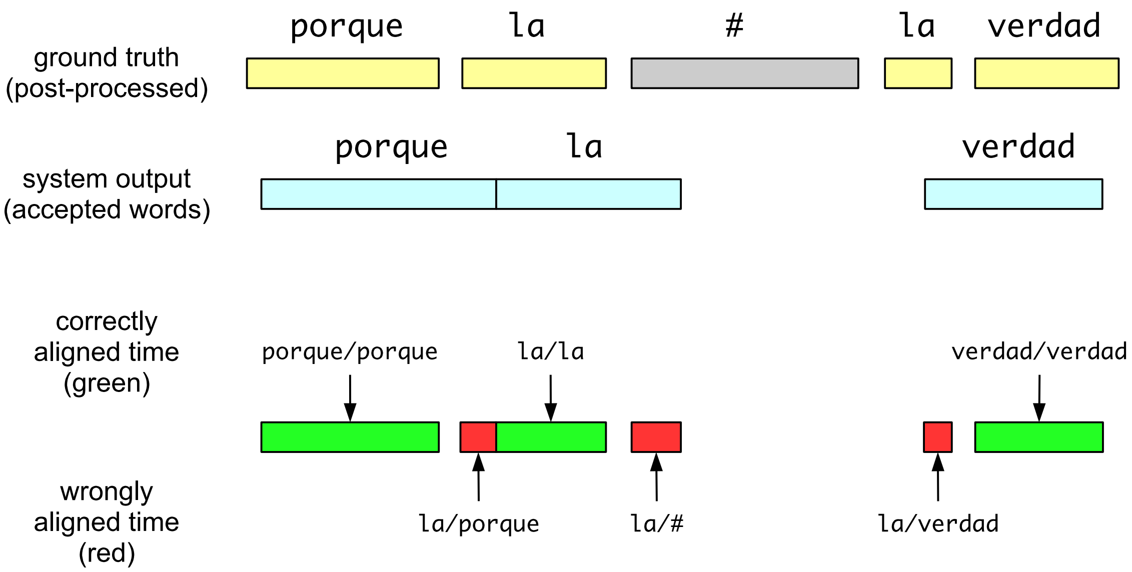
ground truth (post-processed)
porque
la
#
la
verdad
system output (accepted words)
porque
la
verdad
correctly aligned time (green)
porque/porque
la/la
verdad/verdad
wrongly aligned time (red)
la/porque
la/#
la/verdad

Figure 2. The output of the text-to-speech alignment system is aligned to the ground truth to get the sequence of sub-segments which is used to compute the performance metric.

The participants should take into account that the higher the number of accepted segments, the higher the potential amount of correctly aligned speech, but also the higher the risk of having a large amount of wrongly transcribed speech. To find the optimal balance between both events, a suitable confidence score threshold should be applied to make decisions. The scoring script provided by the organization explores all the possible thresholds that can be applied to make decisions, and outputs the optimal score and threshold.

The scoring script

A scoring script and a ground truth file are provided along with the development data, so that the participants can evaluate their progress in the proposed task. The script will require a basic installation of Python 3 including the `matplotlib` module (used to produce a graphical analysis of system scores). The script must be run on the command line, taking the system alignment output and the ground truth files as input. By default, the collar time is 0.0 seconds. Remind that in this evaluation we are applying a collar time of 0.02 seconds. For instance:

```
./eval_align.py -a align_system.txt -t gt.txt -o out.txt -g out.png -c 0.02
```

A help option (`-h --help`) can be invoked to obtain a message describing how to use the script:

```
./eval_align.py -h

usage: eval_align.py [-h] --alignment-file ALIGN_FILE --groundtruth-file GT_FILE
[--collar-time COLLAR] [--output TEXT_OUT] [--graph GRAPH_OUT]

optional arguments:
  -h, –help
             show this help message and exit
  --alignment-file ALIGN_FILE, -a ALIGN_FILE
             Alignment file
  --groundtruth-file GT_FILE, -t GT_FILE
             Ground truth file
  --collar-time COLLAR, -c COLLAR
             Collar time (not evaluated) around word borders
  --output TEXT_OUT, -o TEXT_OUT
             Output (text) column-formatted file with the scoring results
  --graph GRAPH_OUT, -g GRAPH_OUT
             Output (graph) file representing the scoring results
```

The text output consists of two lines: the first one shows the performance obtained using system decisions; the second one shows the best performance obtained by applying a threshold on the provided scores to make decisions. By default, the text output is written on the console. The graphical output (a PNG file) is optional. It presents the performance obtained by applying system decisions and the evolution of the correctly aligned time, the wrongly aligned time and the difference between them (that is, the performance metric) by using all the possible thresholds to make decisions. The optimal performance and the corresponding threshold are marked on the performance curve. The figure also includes the total time accepted and rejected by applying different thresholds. Obviously, applying the minimum threshold implies accepting all the words of the paired text, which does not usually yield the best performance, while applying the maximum threshold implies rejecting all the words, meaning a performance of 0. A reasonable criterion to make decisions on the test set would be to apply the optimal threshold found on the development set.

Participation conditions and schedule

Research teams aiming to participate in this evaluation must register by submitting the following information to the contact email (see contact information below):

- Team name
- Institution name and address
- Contact name and email

Once registered, the participants will be given access to the development data and the scoring script. Test data will be released on September 5th, 2022 (see the schedule below). The registered participants commit to submit the output of one or more systems to the evaluation, under the conditions specified above, namely: (1) audio signals cannot be processed directly by human auditors but only by automatic means; and (2) any kind and amount of resources or tools can be applied, provided that they are suitably reported and described in a system description paper. Each participant can submit the output of at most five systems. One of them must be identified as primary, the remaining ones (up to four) being considered as contrastive. For each submitted system, the alignment outputs for both the development and test sets must be included (in separate files, according to the format specified above). Participants will be ranked according to the the alignment performance obtained by their primary system on the test set.

Submissions must be addressed to the contact email (see contact information below) by the established deadline (October 16th, 2022). Each submission should include the output of the developed systems, along with a short description paper. A full paper with the description of the systems and an analysis of the obtained results must be submitted to IberSpeech by October 31st, 2022. The participants commit to present the developed systems and the obtained results at the evaluation workshop that will be held as part of IberSpeech 2022.

This evaluation plan, the accompanying datasets and the scoring script could be further updated in order to fix potential issues, to introduce new conditions, to account for other performance metrics or to extend the analyses. Any change would be emailed to the registered participants and the evaluation plan would be updated at the website of Albayzin 2022 Evaluations.

The schedule of this evaluation will be as follows:

• September 4th, 2022:	Registration deadline
• September 5th, 2022:	Test data released
• October 16th, 2022:	Submission deadline (system outputs + description paper)
• October 18th, 2022:	Performance results submitted to participants
• October 31st, 2022:	Full paper submission deadline
• November 15th, 2022	Albayzin Evaluation Workshop at Iberspeech 2022 (Granada)

Contact information

Luis Javier Rodríguez Fuentes
Grupo de Trabajo en Tecnologías Software (GTTS)
Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología, UPV/EHU
Barrio Sarriena s/n, 48940 Leioa, Spain

email: luisjavier.rodriguez@ehu.eus