Albayzin 2024 Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge
Evaluation Plan (v1, June 10, 2024)

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

luisjavier.rodriguez@ehu.eus

Overview

Automatic speech recognition (ASR) systems are typically designed to process speech signals in a single language. At most, they are able to transcribe a number of foreign words (usually in English) that are commonly used in the target language. However, in bilingual countries such as the Basque Country, people sometimes switch from language to language (a phenomenon known as c*ode switching*), not only when talking with friends or relatives, but also in more formal situations. That is the case of Basque Parliament plenary sessions, where speakers frequently switch from Basque to Spanish (and vice versa) during their turns. Under these circumstances, an ASR system must be able to deal with code switchings and produce bilingual transcripts. This can be done in several ways, depending on the characteristics of the involved languages. The most straightforward method consists on continuously detecting the spoken language and then applying the corresponding monolingual ASR system. In the last years, efforts have been made to integrate this approach into a single ASR system, robust to code switchings and able to transcribe speech in both languages.

The Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge has been specifically designed to compare different approaches to the task of transcribing speech in two different languages, including the simple approach mentioned above but also other more innovative approaches. We hope that the challenge motivates researchers worldwide to further develop ASR technology able to deal with two languages even in the presence of code switchings.

We provide two training and tuning datasets with ground-truth transcriptions which will allow the development and evaluation of ASR technology under the described conditions. The datasets consist of short (3-10 second long) utterances with the corresponding transcriptions, extracted from Basque Parliament plenary sessions. Each utterance might contain speech in a single language (Basque or Spanish) or speech in both languages (thus featuring a code switching event). The transcriptions of the training set are always close but may slightly differ from audio contents, while the transcriptions of the tuning set have been supervised and validated by human auditors so they exactly match audio contents and can be safely used as benchmark to test ASR performance.

The teams participating in this challenge may develop either an integrated bilingual ASR system or two monolingual ASR systems working under the decisions made by a language detection module. However, we encourage them to develop a single integrated (bilingual) ASR system. To that end, we also provide a baseline ASR system which operates in that way. Participants may take this baseline as a starting point to further refine the approach and hopefully improve its performance.

The development phase will span three months (June-August 2024) and should be employed to build and tune the ASR systems. The evaluation (eval) dataset will be released on September 2nd, 2024 and will consist of a new set of speech utterances, also extracted from Basque Parliament plenary sessions. Participants must submit their output files by October 18th, 2024. Each output file will consist of a UTF-8 encoded text file including one line per eval utterance, containing the utterance name followed by the recognized transcription. The output transcriptions for one primary system are mandatory. Besides, the output transcriptions for any number of contrastive systems can be submitted. Teams will be ranked according to the global Word Error Rate (WER) performance of their primary systems on the eval dataset. Each participant will receive their performance results by October 31st, 2024, with no information about other participants' results.

The task

The task consists on automatically transcribing a short input utterance, which is expected to contain speech in Spanish, in Basque or in both languages. As commonly stated in this kind of challenges, it is strictly forbidden to listen to the audio contents or to hire crowdsourcing (human supported) transcription services. An ASR system, or a set of ASR systems, along with any number of auxiliary subsystems, must be applied to automatically get the transcriptions of test utterances. This means that third-party systems (provided that they work without human intervention) can be used. On the other hand, besides the speech and text materials provided specifically for this challenge, any other training or tuning materials can be used. There is 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, as far as possible, provide links to papers, data and/or software repositories that make it easier to reproduce their approach.

Datasets

For the development of ASR systems, two datasets are provided: training and tuning, both containing short (3-10 second long) speech utterances extracted from Basque Parliament plenary sessions (see Table 1). The transcriptions of the training set might slightly differ from audio contents, while those of the tuning set have been strictly supervised by human auditors in order to be safely used as an ASR benchmark. Two training sets are provided: (1) train, which includes all the available training utterances, no matter the quality attributed to their transcriptions; and (2) train-clean, which includes only the most reliable training utterances (those with phone recognition rate higher than 95%). On the other hand, the tuning dataset is further divided into two subsets: dev and test, designed for tuning and testing purposes, respectively. Each dataset is accompanied by an index file that offers comprehensive information on each utterance (one per line): audio filename, language and speaker tags, phone recognition rate (which loosely reflects how close the transcription is to audio contents), utterance length (in seconds) and transcription.

Table 1. *Duration (in hours) of the training and tuning datasets provided for this challenge.*

Set	Subset	Total	Spanish	Basque	Bilingual
Training	**train**	1445.1	1018.6	409.5	17.0
	train-clean	1315.5	937.7	363.6	14.2
Tuning	**dev**	7.6	4.7	2.6	0.3
	test	9.6	6.4	2.8	0.4

It must be noted that, while the distribution of speakers is balanced in terms of gender, the distribution of languages is not balanced, with Spanish and Basque making up approximately 70% and 30% of the datasets, respectively. Since language tags are associated with both training and tuning utterances, monolingual ASR systems could be developed if desired. Also, language tags could be used to train a Basque-Spanish language detector. Note that, since most of the speakers contribute data to both the training and tuning sets, models will be strongly adapted to those speakers and ASR performance figures will be better than could be expected under strict speaker independence conditions.

To rank the ASR systems developed for this challenge, an independent set of utterances will be released: the *eval dataset*, of about the same size as the tuning dataset. In this case, the index file distributed to participants will provide only audio filenames, without any further information. Since eval utterances will be also extracted from Basque Parliament plenary sessions, the acoustic conditions, the set of speakers and the distribution of languages will be almost identical to those of the training and tuning sets. Of course, words not seen in training and tuning transcriptions may appear in eval utterances. The ground-truth transcriptions of the eval set will be made available to participants when performance results are submitted to them by October 31, 2024.

Audio data

Speech utterances were originally stored in 16 kHz 16-bit signed single-channel PCM WAV files and then compressed into MP3 files. Audio recordings were made through the audio system (desktop microphones) of the Basque Parliament, thus featuring high SNRs. Generally, each utterance contains speech from a single speaker, in a single language (Basque or Spanish). However, some utterances feature a code switching event and contain speech in both languages. Exceptionally, utterances might contain speech from two speakers, one of which would commonly be the president of the Basque Parliament.

Text data

The only text resources provided for this challenge are utterance transcriptions, encoded in UTF-8 text files (the index files). These texts could be used to derive a vocabulariy, that is, the set of words that can be output by the ASR system, and/or a language model. Note, however, that, depending on the approach, the ASR sytem might not require any vocabulary and would rely only on the set of graphemes used in the transcriptions (including blanks). Similarly, the language model, if used, might not involve sequences of words but sequences of graphemes.

Baseline system

A baseline system, able to output bilingual transcriptions of the input utterances, has been developed on the training set and is freely available to participants. We provide recipes to check its performance on the tuning set, using the dev subset for tuning the system and the test subset for measuring its performance. The baseline system uses an acoustic front-end based on a pre-trained Wav2Vec 2.0 speech encoder, which produces a sequence of frame-level acoustic representations (embeddings). Then, a Connectionist Temporal Classification (CTC) backend process that sequence and outputs a vector of grapheme posteriors for each input embedding. Finally, possibly constrained by the phonological and syntactic restrictions introduced by lexical and language models, a search is performed on the sequence of posteriors to output the sequence of graphemes (including blanks) that maximizes the joint acoustic and syntactic likelihood. This baseline system has attained a global WER of about 3% on the test subset of the tuning dataset.

Performance metrics

Global Word Error Rate (WER) will be used as primary metric to rank ASR systems. The submitted transcriptions will be optimally aligned with the ground-truth transcriptions at the word level. The aggregated number of deletions (D), insertions (I), substitutions (S) and matches (M) derived from such alignments will be used to obtain the global WER, as follows:

$$\mathrm{WER} = \frac{D + I + S}{D + S + M} \qquad (1)$$

where the numerator accounts for the aggregated number of errors and the denominator accounts for the aggregated length of ground-truth transcriptions. We will also report, as a secondary metric, the average WER per utterance, defined as follows:

$$\overline{\mathrm{WER}}_{\mathrm{utt}} = \frac{1}{N} \sum_{k=1}^{N} \frac{D_k + I_k + S_k}{D_k + S_k + M_k} \qquad (2)$$

where N stands for the number of test utterances and D_k, I_k, S_k and M_k stand for the number of deletions, insertions, substitutions and matches found in the optimal alignment of test utterance k, respectively.

Submissions

The output of an ASR system must be packed into a UTF-8 encoded text file containing a line per test utterance, each line including the name of the audio file (as provided in the index file of the eval set) followed by the recognized transcription. For instance:

```
file0723.mp3 a lo que nuestro partido se negó por ser inconstitucional
file1216.mp3 zure egiteak eta zuen esateak ez datoz bat eta
file1134.mp3 erdibideko zuzenketa ez da onartu y por no tener no tienen ni un plan
file0042.mp3 y en este momento tenemos ochenta y cinco mil trabajadores
```

The output file should be named according to the following template:

```
<team>_<system>_<p|c1|c2|…|cK>[_late].txt
```

where <team> is a string (without underscores) identifying the participating team, <system> is a string (without underscores) identifying the ASR system, p is used to denote the primary system while c1, c2, …, cK are used to denote the contrastive systems; finally, the _late suffix will be employed only for submissions made after the established deadline.

Participation rules

Registration

Research teams aiming to participate in this evaluation must register through the website of Albayzin 2024 Evaluations (https://catedrartve.unizar.es/albayzin2024.html), providing the following information:

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

It is strongly recommended that, besides completing the official registration, teams wishing to participate in this challenge contact directly with the organizers (see contact info below).

Downloading the datasets and the baseline system

Once registered, the participants can download the training and tuning datasets and the baseline system through the following *HuggingFace* links:

- Datasets: https://huggingface.co/datasets/gttsehu/Albayzin-2024-BBS-S2T
- Baseline system: https://huggingface.co/gttsehu/wav2vec2-xls-r-300m-bp1-es_eu

Further details about the creation of the datasets and the development and evaluation of the bilingual ASR system provided as baseline in this challenge can be found in the following open access paper:

> *Varona, A.; Penagarikano, M.; Bordel, G.; Rodriguez-Fuentes, L.J. A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology. Appl. Sci. 2024, 14, 1951. https://doi.org/10.3390/app14051951*

Participation conditions

The registered participants commit to submit by the established deadline (October 18, 2024) the output of their ASR systems when applied to the eval dataset, along with a description paper with enough details about the developed systems and the results obtained during the development process. The eval dataset will be released through a *HuggingFace* link which will be submitted to the registered participants by September 2, 2024. System development and data processing must be done under the two following conditions: (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 the system description paper.

There are two possible modalities of participation in this challenge, depending on the paper format and the kind of presentation at the Albayzin 2024 evaluation workshop:

1. The first participation modality requires formatting the description paper according to the IberSpeech 2024 paper submission template (https://iberspeech.tech/). In this way, the submitted paper will appear in the IberSpeech 2024 proceedings and the

participants will have the opportunity to submit an extended version of their paper to a journal. Moreover, this participation modality implies the commitment to send one or more team representatives to present the developed systems and results at the Albayzin 2024 evaluation workshop, to be held in Aveiro, Portugal in November 2024.

2. The second participation modality does not require any particular format for the description paper, but this paper will not appear in the IberSpeech 2024 proceedings. In this case, participants are allowed to present their systems without physically attending the Albayzin 2024 evaluation workshop. Instead, they could either send a short (5 minute) video explaining the submitted systems and the obtained results, or connect via streaming during the conference to make the presentation. The streaming option is preferred because it would allow for a Q&A session.

Each team can submit the output for any number of systems (at least one). One of them must be identified as primary, the remaining ones being identified as contrastive. For each developed system, a UTF-8 encoded text file with the transcription of test utterances must be submitted, according to the format specified above. Teams will be ranked according to the global WER performance (Equation (1)) obtained by their primary system on the eval dataset. Late submissions will be considered and scored as regular ontime submissions but will not be taken into account to rank teams.

Submission procedure

Submissions must be addressed to the contact email (see contact information below) by the established deadline (October 18th, 2024). Each submission should include the output of the developed systems along with the description paper.

Evaluation plan updates

This evaluation plan could be further updated in order to fix potential issues, to introduce new conditions, to account for other performance metrics, etc. Any change would be emailed to the registered participants and the evaluation plan would be updated at the website of the Albayzin 2024 Evaluations.

Schedule

- July 31, 2024: Registration deadline
- September 2, 2024: Evaluation data is released
- October 18, 2024: Submission deadline (system outputs + description paper)
- October 31, 2024: Performance results are submitted to participants
- November 12, 2024 Official results are presented publicly and published
- November 12, 2024 Albayzin Evaluation Workshop at Iberspeech 2024 (Aveiro)

Contact information

Luis Javier Rodríguez Fuentes
Software Technologies Working Group
Department of Electricity and Electronics (ZTF-FCT)
University of the Basque Country (UPV/EHU)
Barrio Sarriena s/n
48940 Leioa - SPAIN

web: https://gtts.ehu.es
email: luisjavier.rodriguez@ehu.eus