IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Albayzin 2024 Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge: Datasets, Systems and Results

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

Departamento de Electricidad y Electrónica, Facultad de Ciencia y Tecnología, UPV/EHU
Barrio Sarriena, Leioa, Spain

mikel.penagarikano@ehu.eus, amparo.varona@ehu.eus, german.bordel@ehu.eus, luisjavier.rodriguez@ehu.eus

Abstract

Automatic speech recognition (ASR) systems are commonly designed to process and transcribe speech signals in a single language. At most, they can recognize and output a handful of foreign words. However, speech signals sometimes involve two or more languages mixed continuously, including a sizeable amount of code-switching events, where speakers naturally switch from one language to another. Though efforts have been made in recent years to deal with code-switched speech, there is a lack of benchmarks that allow researchers to evaluate ASR systems on this kind of speech. The Albayzin 2024 Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge aims to provide a benchmark for the Basque-Spanish pair of languages, including the datasets needed to build and evaluate bilingual ASR systems. This paper presents the task, the datasets, the challenge conditions, and a publicly available baseline system with a Word Error Rate (WER) of around 3%. Then, the submitted systems are briefly described and their performance is analyzed. The best-performing system achieved a WER of 1.89% on the evaluation set, which means a relative reduction in WER of 39% compared to the baseline system.

Index Terms: automatic speech recognition, Basque, Spanish, code switching

1. Introduction

Automatic speech recognition (ASR) systems are typically designed to process and transcribe speech signals in a single language. At most, they may be able to output a reduced number of foreign words (usually in English) that are employed in the target language. However, in bilingual countries such as the Basque Country, people sometimes switch from one language to the other, a phenomenon known as *code switching* [1], 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 is to continuously detect the spoken language and then apply the corresponding monolingual ASR system. However, in the last years efforts have been made to integrate this approach into a single ASR system, which would be robust to code switchings and able to transcribe speech in several languages [2, 3, 4, 5, 6, 7, 8, 9, 10].

The Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge has been specifically designed to compare the performance of state-of-the-art ASR technology on the task of transcribing speech in two different languages, including the simple approach mentioned above but also other more innovative approaches. The challenge provides training, tuning and evaluation datasets with ground-truth transcriptions. The datasets consist of short (3-10 second long) utterances extracted from Basque Parliament plenary sessions, which may contain speech in Basque and/or Spanish.

The challenge is agnostic with regard to the ASR approach employed by participants, who may develop either an integrated bilingual ASR system or two monolingual ASR systems working under the decisions made by a language detection module. However, teams have been encouraged to develop a single integrated (bilingual) ASR system. To make the task easier, a fully bilingual baseline ASR system with an overall Word Error Rate (WER) of around 3% has been made available to participants, who could take it as a starting point to further refine the approach and hopefully improve its performance.

The challenge included two phases: development and evaluation. During the development phase, which spanned three months (June-August 2024), the participants could build and tune their ASR systems, based on the training and tuning datasets, which included ground-truth information (utterance transcriptions, speaker and language tags). The evaluation phase started on September 10th, 2024, with the release of the evaluation dataset, which consisted of an independent set of speech utterances, also extracted from Basque Parliament plenary sessions, but without any ground-truth information. The deadline for submitting the output of ASR systems was October 18th, 2024. Each output file should consist of a UTF-8 encoded text file with one line per utterance, containing the utterance name followed by the recognized transcription. Each participant had to submit the output of one primary system and may also submit the outputs of any number of contrastive systems. Submissions were ranked according to the overall WER performance obtained on the evaluation dataset. Finally, on October 25th, 2024, the ground-truth of the evaluation dataset was released and participants were informed about their systems' performance. The evaluation plan provides complete information about the BBS-S2T Challenge [11].

2. The task

The task consisted on automatically transcribing a short input utterance, which was expected to contain speech in Spanish and/or Basque. It was 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, should be applied to automatically get the transcriptions of test utterances. This

10.21437/IberSPEECH.2024-67

means that third-party systems could be used, provided that they worked without human intervention. On the other hand, besides the speech and text materials provided specifically for this challenge, any other training or tuning materials could be used. There was no limit to the type or amount of resources that the participants could use to perform the task, as long as they described the employed methods and resources with enough detail and, as far as possible, provided links to papers, data and/or software repositories that make it easier to reproduce their approach.

3. Datasets

For the development and evaluation of ASR systems, three datasets are provided: training, tuning and evaluation, containing short (3-10 second long) speech utterances extracted from Basque Parliament plenary sessions (see Table 1). The transcriptions of the training set may slightly differ from audio contents, while those of the tuning and evaluation sets have been strictly supervised by human auditors in order to be safely used as an ASR benchmark. Two training sets are defined: (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. On the other hand, the tuning set is further divided into two subsets: *tuning-dev* and *tuning-test*, designed for tuning and testing purposes, respectively. The evaluation set (which is about 40% larger than the tuning set) is designed to rank the ASR systems developed for this challenge. Note that, since utterances in the evaluation set have been also extracted from Basque Parliament plenary sessions, the acoustic conditions, the set of speakers and the distribution of languages are almost identical to those of the training and tuning sets. Mismatches may arise only from words not seen in the training and tuning sets. The training and tuning sets can be accessed through a HugginFace repository[1]. The evaluation set can be accessed through a different repository[2].

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 actual audio contents), utterance length (in seconds) and transcription. In the case of the evaluation dataset, the index file provided initially contained only audio filenames, without any further information; the full index file (containing ground-truth transcriptions, speaker and language tags) was distributed to participants one week after the submission deadline, along with the results obtained by their systems.

Table 1: *Duration (in hours) of the datasets provided for this challenge, disaggregated per language.*

Set	All	Spanish	Basque	Bilingual
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
tuning-test	9.6	6.4	2.8	0.4
eval	23.8	15.5	7.3	1.1

[1] https://huggingface.co/datasets/gttsehu/Albayzin-2024-BBS-S2T

[2] https://huggingface.co/datasets/gttsehu/Albayzin-2024-BBS-S2T-eval

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 all 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.

4. Performance metrics

The overall Word Error Rate (WER) will be used as the 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 overall WER, as follows:

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

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:

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

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.

Finally, since some state-of-the-art approaches output sequences of characters (not words), the overall Character Error Rate (CER) and the average Character Error Rate per utterance ($\mathrm{CER}_{\mathrm{utt}}$) were also computed to study how well the recognized sequences of characters matched the ground-truth. CER and $\mathrm{CER}_{\mathrm{utt}}$ are defined analogously to WER and $\mathrm{WER}_{\mathrm{utt}}$, respectively.

5. Baseline system

A baseline bilingual ASR system has been developed on the training set and is freely available to participants [12]. We provide recipes to check its performance on the tuning set, using the *tuning-dev* subset for tuning the system and the *tuning-test* subset for measuring its performance[3]. The baseline system uses a pre-trained Wav2Vec 2.0 speech encoder [13] as front-end, which produces a sequence of frame-level acoustic representations. Then, a Connectionist Temporal Classification (CTC) neural network backend [14] processes that sequence and outputs a vector of grapheme posteriors for each input embedding. This CTC backend is trained on in-domain data from the Basque Parliament (the *train-clean* subset). 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 [15]. This baseline system has attained an overall WER of about 3% on the *tuning-test* subset [12].

[3] https://huggingface.co/gttsehu/wav2vec2-xls-r-300m-bp1-es_eu

6. Systems

Three teams registered to the BBS-S2T Challenge and submitted 11 systems overall (see Table 2). In the following subsections, we summarize the main features of the systems developed for this challenge.

Table 2: *Teams submitting systems to the BBS-S2T Challenge.*

Acronym	Institution	#Systems
Vicomtech	Vicomtech	7
HiTZ-AhoLab	EHU	2
PRHLT	UPV	2

6.1. Vicomtech systems

The Vicomtech team submitted one primary and six contrastive systems, all of them trained on about 5185 hours of training data (3540, 1631 and 14 hours of Spanish, Basque and bilingual speech, respectively), which is nearly 4 times the size of the *train-clean* subset provided by the organization (which is included in the bundle). The submitted systems exploit different frameworks for the development of acoustic models based on neural networks (K2/Icefall, Nvidia's Parakeet, Kaldi and Whisper), in some cases using a Language Model (n-grams or Large Language Models (LLM)) for decoding [16].

The primary system and the first, second and fourth contrastive systems were based on RNN-Transducer (RNN-T) End-to-End models developed with K2/Icefall [17], using 80-dimensional log Mel-filter banks as input, a Zipformer acoustic encoder and a stateless prediction network using Byte Pair Encoding (BPE), and applying standard data augmentation techniques (SpecAugment and speed perturbation).

The primary and second contrastive systems fused the outputs of the two RNN-T models yielding the lowest error rates on the *tuning-test* subset during the training process, using a voting approach and a Llama 3 LLM to generate the best hypotheses. The primary system applied a few-shot prompting strategy by using manually selected in-domain examples to tune the LLM, while the second contrastive system applied a zero-shot prompting strategy (that is, not using any in-domain examples to tune the LLM).

The first and fourth contrastive systems used the best-performing Zipformer-based RNN-T models for offline and online/streaming applications, respectively. They differ only in the type of convolutions allowed in the encoder layer: the first contrastive system allowed both causal and non-causal convolutions while the fourth contrastive system allowed only strictly causal convolutions.

The third contrastive system applied a Fast-Conformer Transducer model based on Nvidia's Parakeet-RNNT-1.1B architecture [18], trained from scratch in a multitask setup using a Transducer decoder with RNN-T loss and a BPE tokenizer on top of the decoder.

The fifth contrastive system was based on a Multistream Convolutional Neural Network developed by means of the Kaldi toolkit. The input to the network was a sequence of 40-dimensional MFCC coefficients augmented with speed and volume perturbations, to which 100-dimensional ivectors were appended. For decoding, a 3-gram model was trained based on in-domain and generic texts in Basque and Spanish.

The sixth contrastive system exploited the well-known Whisper ASR engine [19]. The 1.55B large-v3 version of the Whisper model, originally trained on more than 5 million hours of multilingual weakly labeled speech, was fine-tuned using Low-Rank Adaptation (LoRA) [20] with all the available data in Basque and Spanish. Since Whisper performs single-language ASR, a language detection module was trained on the datasets provided for the challenge and was applied in sliding windows of 3 seconds with 1-second time shifts.

6.2. HiTZ-AhoLab systems

The HiTZ-AhoLab team submitted two (primary and contrastive) systems to the challenge. Both systems, developed under the Nvidia NeMo framework, were based on a Conformer-Transducer model [21], involving a RNN-T decoder with subword-level (BPE-based) encodings. Language model rescoring was applied using a combination of n-grams and either greedy or modified Adaptive Expansion Search (mAES) decoding [22].

The acoustic models for this challenge were obtained by fine-tuning pretrained models. Firstly a Conformer-Transducer model was trained using 1340 hours of out-of-domain speech data in Spanish. Then, this model was fine-tuned on 548 hours of out-of-domain speech data in Basque. Finally, the model was fine-tuned on in-domain speech data using the *train-clean* subset. The language models were trained on the transcriptions of the *train-clean* subset. The hyperparameters were optimized on the *tuning-dev* subset and performance was evaluated on the *tuning-test* subset.

After preliminary experiments to explore different configurations, the submitted systems were set to use 128 subwords, the primary system implementing mAES decoding and the contrastive system using greedy decoding. More details can be found in [23].

6.3. PRHLT systems

The PRHLT team submitted two (primary and contrastive) systems to the challenge, putting the focus on real-world application, that is, on reducing the computational footprint (memory, training and inference times) of the approach, rather than on performance. The two PRHLT systems were developed from scratch using only the datasets provided for this challenge. Both systems employed the same non-autoregressive end-to-end architecture: a Branchformer-based Mask-CTC model [24, 25], implemented by means of the open-source ESPNet toolkit.

The primary system, with 43.3M parameters, implements the whole architecture, while the contrastive system removes the masking layer and the CMLM decoder, reducing the model size to 33.7M parameters. More details can be found in [26].

7. Results

Table 3 presents the performance achieved by the submitted systems on the *tuning-test* subset (as reported by participants) and on the *evaluation* subset. Results for the baseline system are included too for reference, using two variants: the first performs an unconstrained search on the sequence of posteriors provided by the CTC backend while the second applies a 3-gram language model (trained on bilingual in-domain texts extracted from *train-clean* transcriptions) to constrain the search.

The best performance on the evaluation set was obtained by the primary system of Vicomtech, with a WER of 1.89%, meaning a 39% relative error reduction with regard to the unconstrained baseline system. In fact, the primary, first and second

Table 3: *Performance attained by systems submitted to the BBS-S2T Challenge on the* tuning-test *set (as reported by participants) and on the* evaluation *set. Performance figures for two variants of the baseline system are included too for reference.*

	tuning-test	*evaluation*			
System	%WER	%WER	$\%\mathrm{WER}_{\mathrm{utt}}$	%CER	$\%\mathrm{CER}_{\mathrm{utt}}$
Baseline	3.34	3.11	3.64	1.01	1.17
Baseline-3gram	2.95	2.65	3.16	0.97	1.13
Vicomtech-p	2.34	**1.89**	2.26	0.79	0.94
Vicomtech-c1	2.35	1.91	2.27	0.80	0.94
Vicomtech-c2	2.34	1.90	2.26	0.79	0.93
Vicomtech-c3	2.88	2.64	3.19	1.13	1.36
Vicomtech-c4	2.89	2.60	3.08	1.08	1.27
Vicomtech-c5	3.89	3.59	4.15	1.15	1.32
Vicomtech-c6	4.02	4.13	4.88	1.44	1.63
HiTZ-AhoLab-p	2.67	2.15	2.57	0.85	1.00
HiTZ-AhoLab-c1	2.74	2.15	2.57	0.85	1.00
PRHLT-p	3.40	3.44	3.99	1.14	1.32
PRHLT-c1	3.50	3.56	4.14	1.17	1.35

contrastive systems of Vicomtech showed almost the same performance, revealing that: (1) the few-shot prompting strategy in the LLM did not help compared to the simpler zero-shot strategy; and (2) the voting fusion approach did not help either, compared to using the best single Zipformer-based RNN-T model. On the other hand, the worse performance of Vicomtech-c4 compared to Vicomtech-c1 reveals the importance of using both causal and non-causal convolutions instead of strictly causal convolutions. Finally, Vicomtech-c3 performed slightly worse than Vicomtech-c4, but clearly better than the two remaining systems (Vicomtech-c5 and Vicomtech-c6), so we conclude that RNN-T systems are superior to the other approaches explored by Vicomtech.

The second-best primary system was the one submitted by HiTZ-AhoLab, with 2.15% WER. This approach is similar to Vicomtech-c3 but yielded better performance, probably due to the way pretrained models were created and to the n-gram rescoring method. In this regard, using mAES decoding did not seem to help compared to using greedy decoding, because the HiTZ-AhoLab contrastive system used greedy decoding and performed exactly the same.

The third-best primary system was the one submitted by PRHLT, with 3.44% WER. The PRHLT contrastive system performed even worse, with a WER of 3.56%. Both figures are worse than those of the baseline systems. These results can be explained in several ways: (1) PRHLT systems did not leverage pretrained models, neither for fine-tuning nor for extracting self-supervised acoustic representations; (2) the models were trained from scratch using only the *train-clean* subset provided for the challenge; and (3) no language model was used to condition the search for the optimal sequence of words.

By analyzing the obtained results, best performance seems to depend not only on the approach (RNN-T systems perform better than other types of systems) but also on the amount of speech data used for tuning or training the models: Vicomtech used 5185 hours of training data, HiTZ-AhoLab used 3203 hours overall and PRHLT 1315.5 hours (the *train-clean* subset). On the other hand, the baseline systems strongly relied on a pretrained model (the Wav2Vec 2.0 embedding extractor) trained on thousands of hours of unlabeled speech, the *train-clean* subset being used only to train the CTC backend; besides, the second baseline system used an in-domain 3-gram language model to constrain the output.

With regard to the datasets provided for the challenge, they seem to be consistent, since no overfitting is observed and WER figures are quite similar for the *tuning-test* and *evaluation* subsets, with all systems yielding slightly better results for the latter. The average WER per utterance is remarkably higher than the overall WER in all cases, which tell us about variability in the set of testing utterances. Finally, CER performance is around 1%, ranging from 0.79% (Vicomtech-p) to 1.17% (PRHLT-c1). This means that all systems managed to get high-quality transcriptions, with approximately one erroneous character every 100 decoded characters. Again, the average CER per utterance is remarkably higher than the overall CER in all cases, telling us that errors are not distributed uniformly among the set of testing utterances.

7.1. Performance by language

An interesting line of analysis in a bilingual ASR challenge regards how performance depends on the language. In this challenge, three types of utterances are considered depending on the spoken language: Basque, Spanish and bilingual (code-switched speech). Table 4 presents the WER performance, disaggregated per language, obtained by the primary systems on the *evaluation* set.

Table 4: *Overall WER (%) per language obtained by the primary systems on the* evaluation *set. Baseline systems are included too for reference.*

System	All	Spanish	Basque	Bilingual
Baseline	3.11	2.65	4.57	2.81
Baseline-3gram	2.65	2.19	4.10	2.42
Vicomtech-p	1.89	1.59	2.86	1.66
HiTZ-AhoLab-p	2.15	1.75	3.32	2.59
PRHLT-p	3.44	3.07	4.55	3.42

Clearly, the most difficult utterances are those in Basque. On average, error rates are 1.7 times higher for Basque than for Spanish. The most reasonable explanation is that the datasets used in this evaluation (including those provided by the organization) were strongly unbalanced towards Spanish. More datasets in Basque should be collected in the future to address this imbalance.

Table 5: *Number of parameters of the models, CPU/GPU hardware (Hw), memory (GB) and processing time (real-time factor, xRT) reported by participants for training and decoding. Information for the baseline systems is included too.*

		Training			Decoding		
System	#params	Hw	Mem(GB)	xRT	Hw	Mem(GB)	xRT
Baseline	300M	4xRTX-A5000	120.0	0.029	1xRTX-A5000	5.2	0.005
Baseline-3gram	300M	4xRTX-A5000	120.0	0.029	1xRTX-A5000	5.3	0.006
Vicomtech-p	8B+2x148M	2xL40	5.7	0.166	1xL40+RTX	6.7	0.13
Vicomtech-c1	148M	2xL40	5.7	0.166	1xL40	5.4	0.02
Vicomtech-c2	8B+2x148M	2xL40	5.7	0.166	1xL40+RTX	6.7	0.04
Vicomtech-c3	1.1B	1xL40S	20.1	0.042	1xL40S	3.1	0.65
Vicomtech-c4	148M	3xL40	5.7	0.073	1xL40	5.4	0.02
Vicomtech-c5	20M	5xGTX	5.0	0.287	10xCPU	8.0	3.36
Vicomtech-c6	1.55B	2xA40	12.0	0.048	1xRTX	3.0	0.55
HiTZ-AhoLab-p	119M	NxA100-SXM4	–	0.07	A100-SXM4	–	0.17
HiTZ-AhoLab-c1	119M	NxA100-SXM4	–	0.07	A100-SXM4	–	0.006
PRHLT-p	43.3M	RTX4090	–	0.055	RTX4090	–	0.01
PRHLT-c1	33.7M	RTX4090	–	0.055	Intel-i3-6100	–	0.03

Another interesting observation is the relatively low error rates obtained for code-switched (bilingual) utterances. This may be explained by the dominance of Spanish over Basque in bilingual utterances, so the WER of bilingual utterances is close to that of Spanish utterances. In this regard, the primary HiTZ-AhoLab system was the one that presented the greatest difficulties with bilingual utterances, with an increase in WER of 48% compared to the WER obtained on Spanish utterances.

7.2. Computational issues

Not only the performance is relevant, but also the computational requirements of the proposed approaches, which may be beyond the reach of end users. An ASR system that can be implemented on a relatively weak platform and meets low latency constraints for real-time operation is crucial in real-world applications. Table 5 presents the computational resources employed by the systems submitted to this challenge, as reported by the participants (with some items missing and other items being underspecified). For the sake of completeness, information for the baseline systems is presented too.

The largest models are Vicomtech-p and Vicomtech-c2, both yielding top performance. Interestingly, a much smaller system like Vicomtech-c1 also offers top performance. On the other hand, one of the largest models, the fine-tuned Whisper model (Vicomtech-c6, with 1.55B parameters), yields the worst performance among all the presented systems. Training times depend on the amount of training data and the hardware (GPU/CPU) employed: the real-time factors range from 0.287 (Vicomtech-c5, meaning 1488 hours) to 0.042 (Vicomtech-c3, meaning 220 hours). In the case of the baseline systems, the real-time factor of 0.029 represents about 38 hours, because only the *train-clean* subset was used. The real-time factors reported for decoding range from 0.006 (HiTZ-AhoLab-c1, meaning less than 9 minutes) to 3.36 (Vicomtech-c5, meaning about 80 hours), though we do not know neither how many GPUs were used in the former case nor the type of CPUs used in the latter case. Among the best performing systems, Vicomtech-c1 is the smallest one, with 148M parameters, and also the fastest in decoding, with a real time factor of 0.02 (i.e. about 29 minutes).

8. Summary, conclusions and future work

In this paper, the main features of the Albayzin 2024 Bilingual Basque-Spanish Speech-to-Text (BBS-S2T) Challenge have been presented, including the datasets, the baseline systems, the performance measures, the systems submitted and the obtained results, which were briefly analyzed. This was the first bilingual ASR evaluation proposed as part of the Albayzin evaluation campaigns.

A total of 11 systems were submitted to the challenge by three participating teams. The best system yielded an overall WER of 1.89%, which represents a 39% relative error reduction with regard to the baseline system. In all cases, the worst performance was found for Basque, with error rates 1.7 times higher (on average) than those obtained for Spanish. This is probably due to the smaller amount of training data (speech and text) available for Basque, compared to Spanish. On the other hand, the performance on bilingual utterances was remarkably good, close to that obtained for Spanish, probably due to the dominance of Spanish over Basque in code-switched utterances.

The main conclusion regarding the applied technologies is that RNN-Transducer systems performed better than other approaches. Among the RNN-Transducer systems, the Zipformer-based RNN-T models using Byte Pair Encoding (BPE) obtained the best performance, but the large amount of data used to train these systems could be as important as the model itself when compared to other RNN-T systems trained on smaller amounts of data. There appears to be a high correlation between the amount of training data and the performance achieved.

Future efforts for the advancement of bilingual Basque-Spanish ASR technology should focus on collecting more data for Basque. Also, given the low error rates obtained in this evaluation, more challenging benchmarks should be developed, involving more speakers and more adverse conditions (spontaneous speech, noisy channels and/or environments, etc.). In particular, future benchmarks should remove the speaker dependence intrinsic to this evaluation: the training, tuning and evaluation datasets should feature disjoint sets of speakers. Future benchmarks may also consider the combination of multilingual ASR and machine translation, i.e. the development of systems that produce texts in a target language (Basque, Spanish or other) regardless of the languages spoken in the input speech.

9. References

[1] P. Gardner-Chloros, *Code-Switching*. Cambridge University Press, 2009.

[2] A. Biswas, E. Yılmaz, E. van der Westhuizen, F. de Wet, and T. Niesler, "Code-switched automatic speech recognition in five South African languages," *Computer Speech and Language*, vol. 71, p. 101262, 2022.

[3] C. Zhang, B. Li, T. N. Sainath, T. Strohman, S. Mavandadi, S. Chang, and P. Haghani, "Streaming end-to-end multilingual speech recognition with joint language identification," in *Interspeech 2022, Incheon, Korea, 18-22 September*, 2022, pp. 3223–3227.

[4] O. H. Anidjar, R. Yozevitch, N. Bigon, N. Abdalla, B. Myara, and R. Marbel, "Crossing language identification: Multilingual asr framework based on semantic dataset creation and wav2vec 2.0," *Machine Learning with Applications*, vol. 13, p. 100489, 2023.

[5] K. Dhawan, K. Rekesh, and B. Ginsburg, "Unified model for code-switching speech recognition and language identification based on concatenated tokenizer," in *6th Workshop on Computational Approaches to Linguistic Code-Switching*, Singapore, December 2023, pp. 74–82.

[6] H. Yu, Y. Hu, Y. Qian, M. Jin, L. Liu, S. Liu, Y. Shi, Y. Qian, E. Lin, and M. Zeng, "Code-switching text generation and injection in mandarin-english ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023, Rhodes Island, Greece, June 4-10*, 2023, pp. 1–5.

[7] C. H. Nga, D. Vu, H. H. Luong, C. Huang, and J. Wang, "Cyclic transfer learning for mandarin-english code-switching speech recognition," *IEEE Signal Processing Letters*, vol. 30, pp. 1387–1391, 2023.

[8] J. J. van Vüren and T. Niesler, "Improving Under-Resourced Code-Switched Speech Recognition: Large Pre-trained Models or Architectural Interventions," in *Interspeech*, 2023, pp. 1439–1443.

[9] B. Yan, M. Wiesner, O. Klejch, P. Jyothi, and S. Watanabe, "Towards zero-shot code-switched speech recognition," in *Proc. ICASSP 2023*, May 2023, pp. 1–5.

[10] E. Y. Ugan, N.-Q. Pham, and A. Waibel, "DECM: Evaluating bilingual ASR performance on a code-switching/mixing benchmark," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, May 2024, pp. 4468–4475.

[11] M. Peñagarikano, A. Varona, G. Bordel, and L. J. Rodriguez-Fuentes, *Albayzin 2024 Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge Evaluation Plan*, Spanish Thematic Network on Speech Technologies (RTTH), June 10, 2024, https://catedrartve.unizar.es/reto2024/BBS-S2T2024_Evalplan.pdf.

[12] A. Varona, M. Pengarikano, G. Bordel, and L. J. Rodriguez-Fuentes, "A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology," *Applied Sciences*, vol. 14, no. 5, p. 1951, 2024, https://doi.org/10.3390/app14051951.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *23rd International Conference on Machine Learning*, 2006, p. 369–376.

[15] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[16] J. C. Vásquez-Correa, A. Álvarez, H. Arzelus, S. A. Moreno-Acevedo, A. González-Docasal, and J. M. Martín-Doñas, "The Vicomtech Speech Transcription Systems for the Albayzin 2024 Bilingual Basque-Spanish Speech to Text (BBS-S2T) Challenge," in *Proceedings of IberSPEECH 2024*, Aveiro, Portugal, November 11-13, 2024.

[17] F. Kuang, L. Guo, W. Kang, L. Lin, M. Luo, Z. Yao, and D. Povey, "Pruned rnn-t for fast, memory-efficient asr training," in *Interspeech 2022*, 2022, pp. 2068–2072.

[18] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, and B. Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, 23–29 Jul 2023, pp. 28 492–28 518.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[21] H. Xu, F. Jia, S. Majumdar, S. Watanabe, and B. Ginsburg, "Multi-blank transducers for speech recognition," in *ICASSP 2023*, 2023, pp. 1–5.

[22] J. Kim, Y. Lee, and E. Kim, "Accelerating rnn transducer inference via adaptive expansion search," *IEEE Signal Processing Letters*, vol. 27, pp. 2019–2023, 2020.

[23] A. Herranz, A. García-Sebastián, C. Souganidis, V. García-Romillo, A. Bellanco, E. Navas, I. Hernáez-Rioja, and I. Saratxaga, "HiTZ-AhoLab ASR System for the Albayzin Bilingual Basque-Spanish Speech to Text Challenge," in *Proceedings of IberSPEECH 2024*, Aveiro, Portugal, November 11-13, 2024.

[24] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 17 627–17 643. [Online]. Available: https://proceedings.mlr.press/v162/peng22a.html

[25] Y. Higuchi, S. Watanabe, N. Chen, T. Ogawa, and T. Kobayashi, "Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict," in *Interspeech 2020*, 2020, pp. 3655–3659.

[26] D. Gimeno-Gomez and C.-D. Martinez-Hinarejos, "The PRHLT Speech Recognition System for the Albaizin 2024 Bilingual Basque-Spanish Speech to Text Challenge," in *Proceedings of IberSPEECH 2024*, Aveiro, Portugal, November 11-13, 2024.