Assignment 2: Multilingual Speech Recognition

Winnie Chang
winniech

Yash Mathur
ymathur

We chose ESPnet as the choice of our framework for this assignment. [1] [2]

1 Individual ASR Models

1.1 Guarani

1.1.1 Architecture

We trained three baseline architectures: a Transformer-based model (Vaswani et al., 2017), a Conformer (Gulati et al., 2020) Encoder-based model , and a Branchformer (Peng et al., 2022) Encoder-based model. The Transformer encoder has 12 blocks, Conformer encoder has 8 blocks while the Branchformer encoder used 12 blocks. The decoders for all three baselines had 3 blocks [3].

1.1.2 Training and Hyperparameters

For all 3 baselines, we used the Adam optimizer with a learning rate of 0.0005 and 800 warmup steps. Furthermore, we used joint CTC/attention training with a weight of 0.3.

Since, we are working with an extremely low-resource language, in order to provide the model with more data to train, we perform speech perturbation of 0.9, 1.0, and 1.1 for all experiments.

To empirically evaluate the impact of tokenization schemes on end-task performance, we evaluate word level tokenization and Byte-Pair encoding (Sennrich et al., 2016) with varying vocabulary size.

We train each model for a maximum of 200 epochs with patience set to 10.

1.1.3 Results

Our Transformer, Conformer and Branchformer baseline model(s) achieve a poor CER and WER. On further inspecting the predictions, we observe that the model was repeating tokens in its predictions which provided us intuition that the model was overfitting and/or model was unable to find discernible patterns in the speech signals due to large token lengths(word level tokenization). This led us to explore alternative tokenization schemes such as Byte-Pair encoding. We further experimented with high vocabulary size for BPE (nbpe=500) and low vocabulary size for BPE (nbpe=100) and concluded utilizing a lower BPE is beneficial for improved learning based on our intuition. This is despite the empirical results because we believe a fine-grained tokenization scheme would enable a better mapping between speech signals and tokens. In conclusion, we trained five different baseline models for Guarani. The results for each of the models are summarised in Table 1.

1.1.4 Analysis

The results across models are aligned with our intuition that modeling an ASR system for a low resource language like Guarani has challenges where the model doesn't have access to a rich set of words used in the language. Additionally, it appears the models exhibit strong overfitting on the data which should be a consequence of the limited data as well along with a high model capacity. It is also evident from the results that since the current pipeline doesn't have access to language modeling capabilities it is able to perform better in identifying characters but is unable to form valid/correct words. We further observe a strong sensitivity(minor changes in learning rate leads to high variance in metrics) to changes in hyperparameters signifying the need for hyperparameter tuning for optimal performance.

1.2 Quechua

1.2.1 Architecture

We used the same architecture setups as Guarani for our experiments: transformer, conformer, and branchformer.

[1] Our code is available on GitHub

[2] Our WANDB Report for Guarani is available here and for Quechua here

[3] The entire configuration for all the models trained for the assignment is available as part of the submitted recipes

Model	Trained LM	Tokenization	Augmentation	WER	CER	TER
Transformer	No	Word	No	99.5	91.2	-
Branchformer Encoder	No	Word	No	99.5	91.2	-
Conformer Encoder	No	Word	No	99	93	-
Branchformer Encoder	Yes	BPE-100	No	101.1	164.7	193.1
Branchformer Encoder	Yes	BPE-500	No	104.8	83.6	90.7
Branchformer + Frontend Hubert (Frozen)*	Yes	BPE-100	Yes	57.1	13.5	16.4
Branchformer + Frontend XLS-R (Frozen)*	Yes	BPE-100	Yes	**51.6**	**11.8**	**14.3**
Multilingual Branchformer + Frontend HuBERT	Yes	BPE-100	Yes	66.0	18.6	18.9

Table 1: Validation Set performance metrics for Guarani, * denotes a more shallow decoder being used

Target Sentence	Branchformer	Branchformer + BPE	Branchformer + HuBERT Frontend
upéicha AVEI OPURAHEIJEPéKURI AMBUE TETãME.	HETA	upéicha MBA'E HESE PETEĩ MITãMI.	upéicha avei opurahijepékuri ambO'etãme

Table 2: Exemplar Outputs from Guarani Models

1.2.2 Training, Hyperparameters, and Data

Following our experiments in Guarani, we used the same hyperparameter setup and data augmentation as in 1.1

From our experiments with Guarani and our intuition that sub-word level tokenization would improve ASR by making it easier to map graphemes to sound segments, we observed that though word tokenization was empirically performing better than BPE, BPE makes more sense, so we also trained a branchformer model (which performed the best) with a BPE of 100.

For the specific Quechua dataset we used, the test set did not have aligned text so we took the dev set as the test set and split from the train set the original number of samples in the dev set to use for validation.

1.2.3 Results

Model	WER	CER	TER
transformer	100.0	92.0	-
conformer	99.6	91.3	-
branchformer	99.8	77.4	-
branchformer_BPE100	99.3	85.8	111.4

1.2.4 Analysis

We observe that even with the SOTA branchformer architecture, model performance with respect to WER lags behind CER. A possible reasoning is that Quechua is an agglutinative language and thus Quechuan words tend to be very long. In a later experiment, we attempt to use language modeling to mitigate this issue.

Furthermore, upon inspection of the Quechua dataset we are using, there seems to be a good number of code switching samples where Spanish is mixed in. This in particular could also be confusing the model which now has to deal with "noise" contaminating the samples or modeling two languages at once while having very little data to work off of.

2 Multi-Task, Multilingual Joint Training and Using Pretrained Models

2.1 Guarani

2.1.1 Using Pretrained Models

Based on the observed improvements, we utilize two frontends: HuBERT (Hsu et al., 2021) and XLS-R 128 (Babu et al., 2021) with goal of exploiting better learned representations from a frozen

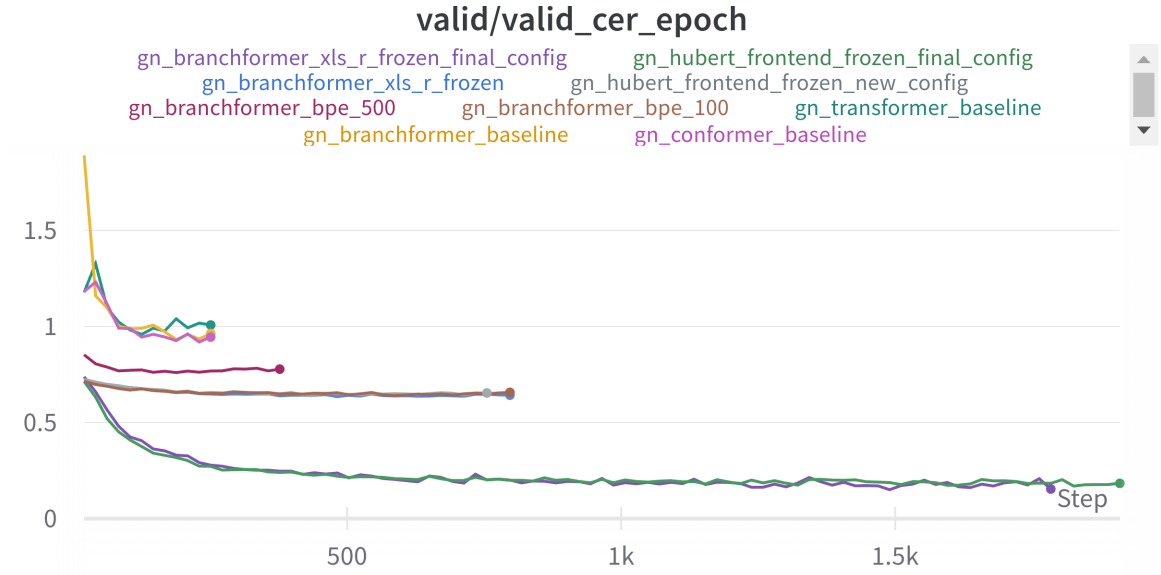

Figure 1: Validation Split CER for Guarani across models

frontend. We combine the frontend with data augmentation techniques such as SpecAugment (Park et al., 2019). We further introduced training a language model to improve the ability of the ASR system to model its word predictions. This was in line with our hypothesis that a smaller vocabulary enables the model to learn mapping between fine-grained sound signals to their corresponding token in the language.

2.1.2 Multi-Task, Multilingual Joint Training

For selecting an auxiliary language to perform multilingual training, we utilize geographic proximity to the regions where Guarani is spoken and choose Argentinian Spanish. The baseline multilingual model follows the same architecture as the one used in Section 1.1. The HuBERT frozen frontend is the second model chosen for multilingual training.

2.1.3 Analysis

We observe a significant improvement while using large pretrained models as a frontend which follows performance improving by using richer representations. We see a significant improvement by utilising a language model which enables the model to make more accurate word/token level predictions based on current input. We further modify the decoder to use reduced number of blocks to improve training times and found that the overall performance improved with this as well. We observe that multilingual training offers marginal improvements to the results from Section 1.1. This could be caused due to not choosing the optimal auxiliary language and/or the limited data (which are expected consequences of working with Low-Resource Languages).

2.2 Quechua

2.2.1 Using Pretrained Models

We followed a similar setup as Guarani for our pretrained model experiments. We utilized both HuBERT and XLS-R 128 as a frozen front-end layer for feature extraction, an LM to improve decoding generation, and SpecAugment to provide additional data augmentation.

We ran 2 experiments, one with a HuBERT frontend and one with a XLS-R 128 front end.

Frontend	WER	CER	TER
HuBERT	97.5	53.5	72.9
XLS-R 128	93	45.3	64.3

We observe that the XLS-R 128-based model outperformed the HuBERT-based by a significant margin. This could be because HuBERT is trained only on English data while XLS-R is trained on 128 languages, thus having a better representation distribution. Furthermore, although WER is still as bad as the baselines, we see CER improve from our baseline models and it is likely to be attributed to the fact that the pretrained models have better feature representations than training from scratch.

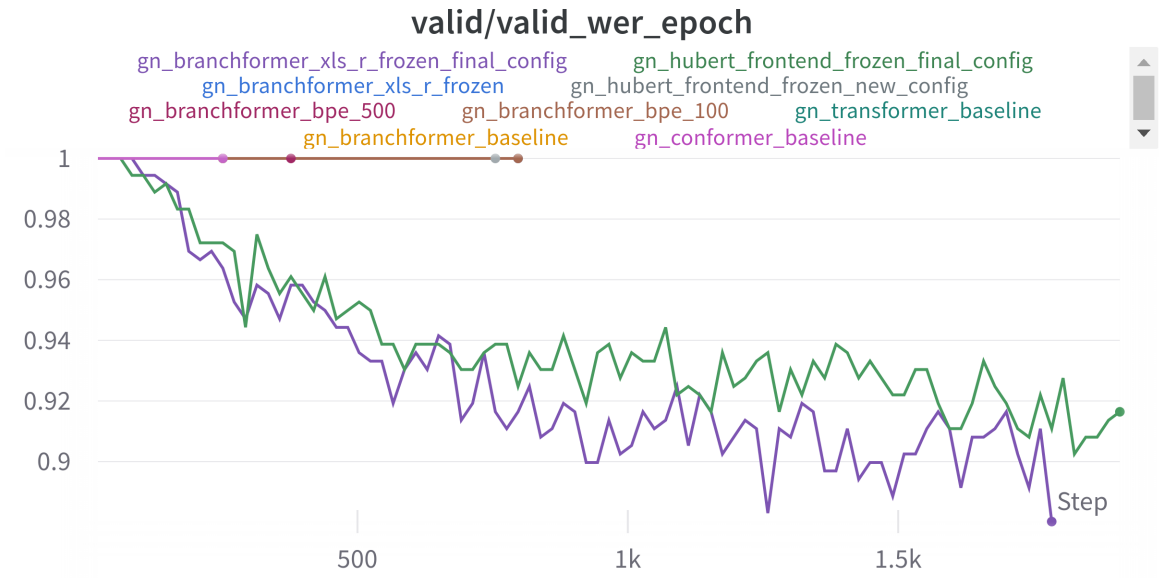

Figure 2: Validation Split WER for Guarani across models

2.2.2 Multi-Task, Multilingual Joint Training

Similar to Guarani, when considering which language to select as an auxiliary language to perform multilingual training, we utilized geographical proximity in addition to the knowledge that Quechua speakers code switch to choose Peruvian Spanish.

We trained 2 different models under this setting, one with the best baseline model from 1.2 and one with the best pretrained model from the previous section. Hyperparameters during training remained the same.

Model	WER	CER	TER
baseline	99.4	86.4	87
XLS-R 128	106.8	47.7	61.6

We observe here that the multilingual model based on the best baseline model did not see improvements from the other baselines. Neither did we see improvements with the model based on XLS-R 128 with the monolingual version. This could be attributed to a couple reasons. First, we could have chosen the wrong auxiliary language. Our survey of which language to choose came at a very high level and perhaps that particular dialect of Spanish is not very influential or similar to Quechua nor the Spanish that is code switched in the data. Perhaps a deeper dive into the linguistic feature of Quechua would be necessary to choose a better auxiliary language. Secondly, it could be due to the lack of data. In an effort not to overwhelm the model with Spanish data, we chose to only take a subset of the Spanish data so that it was roughly a 1-1 ratio. However, this only ends up giving us about 1200 samples of audio total to train with, after speech perturbation.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global

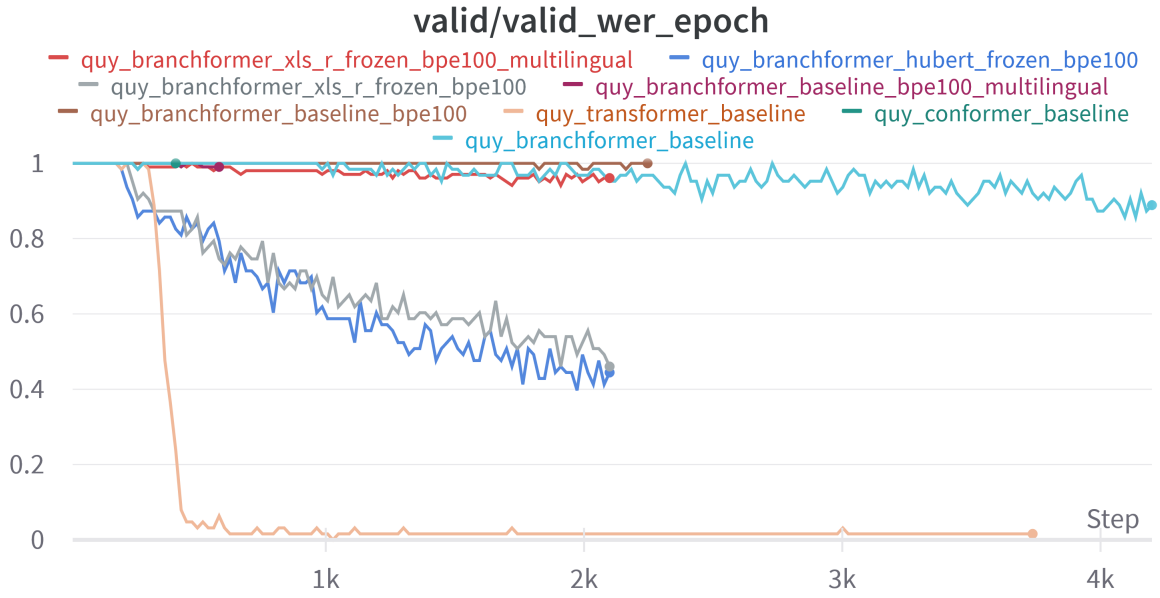

Figure 3: Validation Split WER for Quechua across models

context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

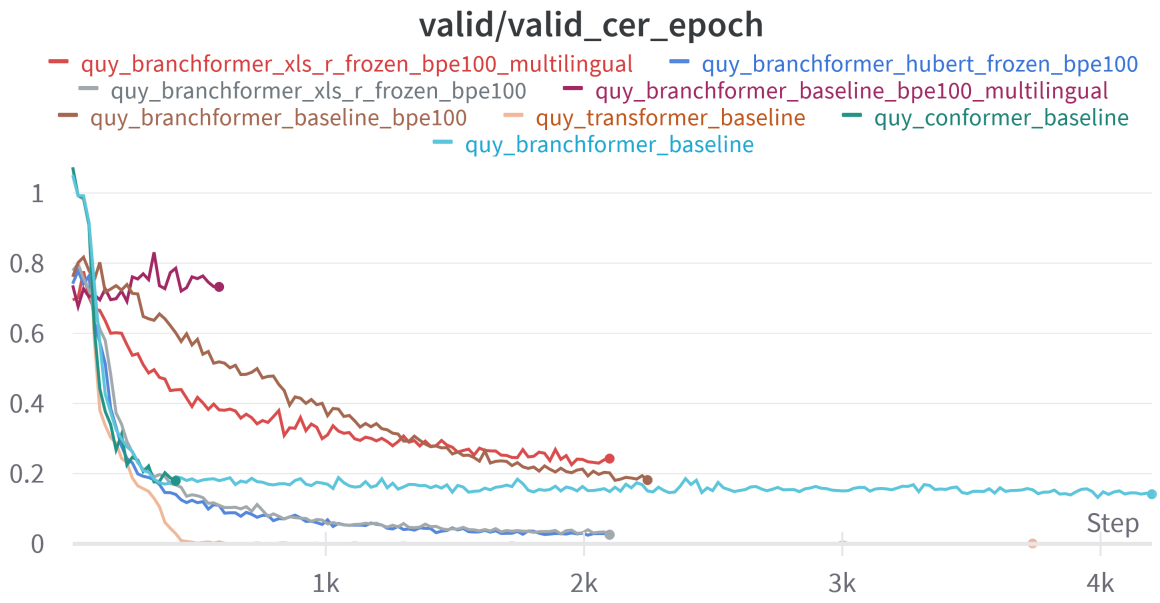

Figure 4: Validation Split CER for Quechua across models