Multilingual Speech Recognition

Yerin Han, Su Park, Malaika Vijay
Carnegie Mellon University Language Technologies Institute
Pittsburgh, PA
yerinh,suminpar,mvijay@cs.cmu.edu

1 Introduction to ASR

This project aims to train an ASR System on a low resource language. ESPnet is an open-source end-to-end speech processing platform that enables end-to-end automatic speech recognition. We use ESPnet to train a variety of ASR models on recordings of African Accented French collected from speakers across various regions of Africa. Our experiments include identifying the optimal Byte Pair Encoding vocabulary size, incorporating language modeling, and employing self supervised representation learning with HuBERT and wav2vec 2.0. The best performing model was a combined CTC and Attention model with a conformer encoder and a transformer decode which resulted in 29.2 WER, 18.7 TER, and 9.0 CER for the test dataset.

2 Dataset Description

We trained an ASR system on the OpenSLR African Accented French Dataset. The original dataset contains roughly 22 hours of both read and conversational speech recordings split across 3 subsets of data - CA16, Yaounde, and Niger. However, we retain only one subset (CA16). The pruned dataset consists of 4000 train, 600 validation and 600 test utterances. While the train and validation set consists of speakers from a single region, Gabon, the test dataset consists of speakers from Gabon (80%), Cameroon (14%), Chad (3%) and Congo (3%). The test set does not contain any recordings from speakers included in the train and validation sets. We preprocess the transcriptions to convert all characters to lowercase and remove all punctuation marks. However, a bug in our code revealed that this removal of punctuation did not get applied to the test and validation sets, resulting in errors related to punctuation.

3 Experiments

We run our experiments using a combined CTC and Attention model with a conformer/transformer encoder and a transformer decoder. We also employed speed perturbation across all experiments.

3.1 Size of Byte Pair Encoding Vocabulary

This set of experiments was aimed at identifying the optimal size of the Byte Pair Encoding (BPE) vocabulary used to tokenize the transcript text. We experimented with 150 and 200 tokens with the conformer encoder and transformer decoder architecture. Our findings indicate that a smaller BPE vocabulary was more suitable to our dataset.

3.2 Language Modeling

Our next set of experiments focused on identifying the effect of language modelling in ASR. In speech recognition, language model and acoustic model are used together to transcribe speech. While the acoustic model transforms the analog sound waves into digital phonemes, the language model limits the hypothetical search space and predicts which word or phrase will follow the current word with what probability. However, we require a sufficiently large corpus of diverse text to train useful language models, which was not a characteristic of the dataset used for these experiments. We test the performance of our ASR system with and without the use of language modelling in the decoder. This set of experiments was carried out with both a conformer and transformer encoder along with a transformer decoder.

3.3 Self-Supervised & Fused

The Hidden-unit BERT (HuBERT) (Hsu et al., 2021) approach for self-supervised speech representation learning uses K-means clustering assignments of masked segments of continuous input to learn both acoustic and language models from continuous inputs to provide aligned target labels.

While HuBERT is related to wav2vec 2.0 (Baevski et al., 2020), the two differ in that the latter uses contrastive loss that requires a meticulous design regarding auxiliary diversity loss. wav2vec 2.0 also requires careful planning around where to sample negative frames and a pre-defined Gumbel-softmax schedule and only focuses on quantizing the waveform encoder output. HuBERT extends upon the aformentioned disadvantages of wav2vec 2.0 by distinguishing the masked prediction representation learning from the acoustic unit discovery step and outperforms it on various fine-tuning scales. For this assignment, we experiment with incorporating HuBERT on its own and fusing it with wav2vec 2.0 to delineate the incremental improvements in performance each brings.

4 Results and Analysis

This section discusses the results of our experiments. Here, our baseline model refers to one with a conformer encoder and a transformer decoder. Our fused model refers to one with a transformer encoder and a transformer decoder with HuBERT and wav2vec 2.0 features.

4.1 Byte Pair Encoding Vocabulary Size

Table 2 shows the results of our experiments on varying BPE vocabulary sizes. Our findings indicate that a smaller BPE vocabulary size leads to better ASR performance on this dataset. We hypothesize that a larger number of BPE tokens naturally results in a larger vocabulary size (more subwords), hence produces a larger embedding matrix used in the decoder and a more complex model. With such little data, more parameters in the network could lead to heavy overfitting which fails to generalize properly on unseen examples.

4.2 Language Modelling

Table 1 shows the results of adding language modelling to our models. While we had expected language modelling to show gains in performance over the baseline model, we find that adding language modeling to our ASR system marginally worsens performance on the test set as seen in tables 2 and 1. We attribute this drop in performance to a poor language model which originates from the absence of sufficient data, both in terms of quantity and variety. The language model is trained on the transcripts of the training data, which consists of text corresponding to only 4000 utterances. Many of these transcriptions are prompts that do not show a large amount of diversity in vocabulary as well.

4.3 Self Supervised Speech Representation - HuBERT & HuBERT + wav2vec 2.0

It's interesting to note that the fusion of wav2vec 2.0 and HuBERT resulted in the worst performance, while LM performed the best across CER, TER, and WER. HuBERT on its own (SSLR) scored quite poorly, yet when combined with Language Modeling, the performance improved by 5 points for WER. One possible explanation is that the strength of LM contributed to an enhanced performance despite the increased model complexity.

HuBERT on its own also outperformed the fused model, as the strength of HuBERT had been validated to surpass that of wav2vec 2.0. The fusion model's relative failure can be attributed to the overfitting that resulted from the already small and regionally imbalanced dataset. If we were given more time, we would like to experiment with just wav2vec 2.0 on its own to evaluate the effect of fusing two self-supervised approaches over just using HuBERT. Future works can also balance the dataset through upsampling or downsampling to validate whether this phenomenon is due to the severe overfitting that cannot be addressed with mere hyper-parameter tuning or that a fusion of two self-supervised learning approaches do not perform well across the board.

4.4 Error Analysis

Manual comparison of the reference text with the generated hypothesis revealed two primary causes of errors. These correspond to the "'" and "-" characters. We erroneously removed these characters during the data preparation stage only from the train set and not from the test and validation set. However, they are an integral part of French morphology (*c'est, voulez-vous*). This resulted in a mere misalignment or omissions being flagged as inaccuracies. Our best ASR model failed on most occurrences of these characters in the reference transcriptions. Figure 1 exhibits some samples of such errors.

5 Limitations

One limitation is the lack of regional variety in the speakers' origins. As the train and validation sets solely included recordings of Gabon-French speakers and the test set was predominantly consisted

	Conformer + LM		Conformer + HuBERT		Conformer + HuBERT + LM		Fused	
Data	**Validation**	**Test**	**Validation**	**Test**	**Validation**	**Test**	**Validation**	**Test**
WER	14.9	30.1	14.0	32	14.1	30.4	18.9	34.9
TER	8.3	19.4	7.6	21.4	7.6	20.8	10.8	22.4
CER	5.3	9.7	4.5	11.5	4.7	11.1	7.3	12.9

Table 1: Results of experiments with Language Modelling and Self Supervised Representations

REF: c ' <space> e s t <space> p r u d e n t <space> d ' <space> a l l e r <space> v o i r <space> u n <space> m é d e c i n
HYP: c * <space> e s t <space> p r u d e n t <space> d * <space> a l l e r <space> v o i r <space> u n <space> m é d e c i n

REF: e s t c e <space> q u e <space> l e <space> m é d e c i n <space> e s t <space> l o i n <space> d ' i c i
HYP: e s t c e <space> q u e <space> l e <space> m é d e c i n <space> e s t <space> l o i n <space> d <SPACE> i c i

REF: p u i s - j e <space> v o i r <space> l a <space> b l e s s u r e
HYP: p u i s * j e <space> v o i r <space> l a <space> b l e s s u r e

REF: a v e z - v o u s <space> u n <space> v é h i c u l e
HYP: a v e z * v o u s <space> u n <space> v é h i c u l e

Figure 1: Error Analysis

	Baseline BPE150		Baseline BPE200	
Data	**Validation**	**Test**	**Validation**	**Test**
WER	15.0	29.2	16.5	34.4
TER	8.5	18.7	11.8	26.1
CER	5.3	9.0	6.5	13.5

Table 2: Performance of Conformer Model with Varying BPE Vocabulary Size

(58%) of Gabon speakers and the remainder were 3 Cameroon, 1 Chad, and 1 Congo speaker, it is hard to extrapolate statistically valid insights regarding how regional differences in accent, syntax, and morphology could have influenced the variation in performances.

6 Conclusion

For this project, we trained an ASR model using ESPnet on a dataset of African Accented French recordings. We studied the effect of BPE vocabulary size, language modelling and self supervised feature representations on performance of the ASR system. Our results, which we personally find somewhat unsatisfactory, nevertheless revealed the risks of applying complex models or fused models to imbalanced, low-resource data which inevitably results in overfitting.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1.