214B Miniproject 2 - Foundation Models for Low-resource ASR

David Castro, [redacted], [redacted]

UCLA

davcstro@g.ucla.edu, [redacted], [redacted]

Abstract

Children speech recognition is an unsolved problem in speech processing which, if improved, could enable the development of applications for classroom environments and education in general. An approach to tackle this ongoing research problem is to make use of self-supervised speech models. In this project, we use the Wav2Vec2 model and the MyST dataset to devise ways of improving WER on children automatic speech recognition. Specifically, this project tackles the small data problem with augmentations including VTLP, pitch variations, and additive noise. Comparing baseline models using the non augmented data set we achieve an improvement of 3.4% on both the test and development sets. Additionally when considering the addition of a language model using 5-gram learned from 10 hours of MyST data, we achieve an improvement of 9.1% on the test set and 9% on the test-dev set from the baseline.

Index Terms: speech recognition, low resource ASR

1. Introduction

Automatic speech recognition (ASR) is the process by which a computer recognizes and understands speech from a user. ASR has improved greatly over the years through the increase in sophistication of neural networks: transformers, language models, etc. A main reason for the fast improvement in ASR is because of researchers open sourcing their neural network architectures such as Wav2Vec [1], Hubert [2], Whisper [3]. Even though ASR's word error rate (WER) has improved there remains many challenges: dialect, accent, and children's speech. The main problem behind these challenges is the amount of data available. In this project Wav2Vec2, the second and improved version of Wav2Vec, will be used to measure WER performance when applied to standard datasets - Librispeech (adult speech) [4] and "My Science Tutor" (MyST - kid speech) [5]. Augmentation of the MyST dataset will be used in order to demonstrate how to improve WER when only a limited dataset is available. In order to further improve the WER, a language model will be introduced. Lastly, hidden representations in foundation models will be discussed and how this may be the next important step in improving ASR.

2. Background

In this project the Librispeech and MyST audio datasets will be used to finetune Wav2Vec2 which has been already pre-trained. Pre-training is the process of feeding large amounts of unlabeled data into the neural network in order for it to learn internal representations of the training data. The pre-trained model can then be finetuned on downstream tasks, in this case ASR. This has been demonstrated to reduce both task-specific training time and the need for large amounts of labeled data. In order to reduce training time, we use the one hour training set for both Librispeech and MyST to finetune the model. The Wav2Vec2 comes with multiple architecture models, each with different layer sizes in the neural network. For this project, we limit finetuning on both the base and large model variations.

Librispeech is a dataset consisting of audiobook recordings while MyST consists of recordings from children interacting with a virtual science tutor. In the majority of cases, children's data is limited in size due to laws protecting the privacy of minors, explaining the limited size of MyST compared to Librispeech. In order to address this, augmentation on the existing data can be done to synthetically increase the size of the training data. For this project, the python package *nlpaug* will be used [6]. There are many augmentation methods; a basic commonly seen augmentation method is adding noise to the existing data to improve recognition for noisy signals. The main method to improve the ASR will be vocal tract length perturbation (VTLP) in order to create new children data from existing adult data. There are many augmentation methods, and each have their own benefits and limitations depending on the situation.

Language models provide an additional source of truth when performing ASR, helping to distinguish between similar sounding words. By using Connectionist Temporal Classification (CTC) decoding, the probability distributions of the next token, generated by the wav2vec2, are further refined when considering the language model. An n-gram language model leverages sequences of *n* consecutive words to assign probabilities to the next words. In order for these language models to best predict the text, they should be trained on text data which approximates the language used in the task.

3. Project Description

The first task of the project aims to set-up an ASR, using wav2vec2 architecture from a huggingface blog [7]. Both the one hour for Librispeech and MyST with a transcription of the speech data are used in order to train the ASR, and then gain a baseline of the WER. Each evaluation for word error rate is done on a separate amount of data from its respective database.

3.1. Data Augmentation

Different augmentations are done on the MyST dataset in order to see what can work in order to improve the ASR.

3.1.1. Spectral Masking

Spectral Masking or masking in augmentation is the act of zeroing out certain portions. In the case of spectral masking, it means to zero out certain frequencies. This allows for the generation of new data because it can remove certain frequencies that may not be as important in the training, or it may remove important frequencies which may possibly worsen the model. Random spectral masking was done on the existing one hour training data being used for children speech. This means each sound file of children's speech in the one hour training set had a different bandwidth of frequencies zero'd out. An example of this can be seen in figure 1.

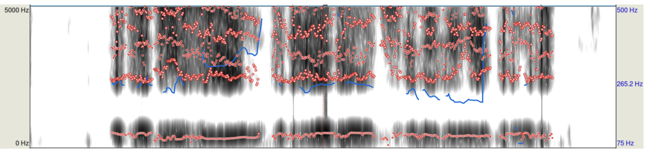

Figure 1: *Frequency Masking Example*

3.1.2. Vocal Tract Length Perturbation

Vocal Tract Length Perturbation (VTLP) aims to shift the formant frequencies of speech. The formants are determined by the length of the vocal tract, hence the name VTLP. Children naturally have higher formant frequencies because their vocal tract length is smaller. The shift is not linear though where formant one and formant two are both higher by the same amount compared to adults, it varies. This makes VTLP very difficult since it takes trial and error or measurements to find the right amount of shifting. In our VTLP, new kid data can be generated by performing VTLP on adult data with a 1.1-1.3x increase to create new "children data".

3.1.3. Pitch

Pitch varies with age and gender. Children naturally have a pitch over 250 Hz while adults range from 100 to 300 Hz depending on gender. In order to improve kid speech, it is necessary to create more data with a pitch over 250 Hz. Existing adult speech from the Librispeech database can have its pitch augmented to be 2-3x higher in order to generate more data for the MyST (kid) ASR.

3.1.4. Noise

Another popular method to artificially create more data is to consider additive noise to existing samples. Here we already begin with children's speech and can then create more "copies" with noise but we need to be careful with how we do this. As we know much of the formants for children speech is often in the higher frequencies for vowels as well as a higher pitch which was discussed in the previous section. Thus the type of noise that we select mustn't contaminate the existing signal to the point that it is no longer a good representative of what a kid would sound like for the utterance. As with the other methods we look to adapt other types of utterances through signal processing approaches and turn them into children speech but here we begin with children speech as being the main differences. In many other scenarios the usage of white noise is often seen as the first option due to its nice properties however this "niceness" doesn't actually work in the context of children speech as it ends up treating all the frequency ranges equally. A more common type of noise in speech recognition domains is then pink noise which is concentrated at lower frequencies and then drops off at a rate of 10 db/decade in the context of Hertz. Thus we can safely apply this to the lower ranges 1-1000 Hz with a lesser effect for higher frequencies where some of the formants may be. As we also know that it is a nice estimator for babble noise it can potentially make systems more robust to different environments especially those in the classroom where these speech recognition systems may be implemented.

3.2. Language Model

A 5-gram language model (LM) will be used in this project. The first learning model was trained on the 10 hours of Librispeech text, and the second was the 10 hours of MyST text. The reason for testing two different models is because this will help us understand the importance of aligning the language model to the task at hand.

Librispeech is a dataset consisting of recordings of adults reading books. This greatly differs from MyST in a few significant ways. First, Librispeech will be biased towards complete and grammatically correct sentences, rather than natural conversation, as MyST is. Second, the language content of Librispeech will be much more diverse than MyST, while the language in MyST will be constrained to science and limited by the vocabulary of children.

Additionally, we also seek to understand how difficult it can be to create text corpora to customize our own language model. To do so, we scraped wikipedia sites by saving the article's body, navigating to a linked article and repeating the process. The starting point for this scraper was a science article, with the intention of aligning the content to the expected language of the MyST dataset.

We use KenLM [8] to train a 5-gram language model on Librispeech, MyST and our custom Wikipedia corpus. We expect to see best results when using the language model trained on MyST.

Table 1: *LibreSpeech Results - 2000 Iterations*

Model	Test-clean	Test-other 3	Dev-clean	Dev-other
Base	0.276	0.368	0.261	0.364
Large	0.169	0.238	0.167	0.234

Table 2: *MyST Results on 1 Hour Training Set - 6000 Iterations*

Model	Augmentations	LM	Test	Test-Dev
Base	None	None	0.394	0.385
Large	None	None	0.338	0.33
Base	Freq-Mask	None	0.384	0.376
Large	Freq-Mask	None	0.352	0.347
Base	VTLP/Pitch	None	0.36	0.351
Base	VTLP/Pitch/Noise	None	0.37	0.358
Base	VTLP/Pitch	LibreSpeech 10Hr.	0.32	0.311
Base	VTLP/Pitch/Noise	LibreSpeech 10Hr.	0.326	0.314
Base	VTLP/Pitch	MyST 10Hr.	0.303	0.295

Table 3: *Learning Model Results for Comparison*

LM-5gram	Test WER
None	0.396
LibreSpeech	0.345
Wikipedia	0.341
MyST	0.326

4. Summary and Discussion

For both Librispeech and MyST, both base models were run in order to see a comparison between the base and large model results. In both cases, an improvement is seen in the WER of the large model. In order to properly train the large model, some trial and error was required for the learning rate in order to see an improvement in accuracy.

Different data augmentations saw different results. Spectral masking had roughly a 1% increase in performance, which does not really show much of improvement. This makes sense because the spectral masking used was random in its bandwidth or frequency selection, meaning it may zero out important frequencies. In order to possibly achieve a higher improvement for WER in spectral masking, further testing would be required with specific frequency bands in order to see what provides the most improvement. VTLP and pitch augmentation was performed on the 1 hour Librispeech dataset in order to see if adult data can be converted to kid data. By adding this newly created data, an improvement of 3.4% was seen. A way of possibly improving the accuracy of VTLP could be separating the female and male speakers in Librispeech and performing more specific value augments to get closer to children's speech. Pink noise augmentation was combined with the VTLP and pitch augmentation described previously. It did not result in an improvement compared to just the VTLP and pitch augmentation since it resulted in a 2.9% improvement from the base model rather than a 3.9%. Noise augmentation is good for situations where you want to improve ASR in noisy situations, but since the dataset for MyST is not in a noisy situation, it may degrade the clean recognition.

Three different 5-gram learning models were generated. Only MyST and Librespeech learning models were used on the trained augmented datasets. An improvement of at least 4% with the Librispeech trained LM compared to without the LM is seen. For the MyST learning model, an improvement of 5.7% is seen than without the LM. The higher improvement with the MyST trained LM is expected because this LM pulls from the same dataset that the evaluation is being done from: children's speech and science subject. The Wikipedia LM is compared with MyST and Librespeech LM on the MyST baseline - 1 hour of data, no augmentations. While Wikipedia performed slightly better than LibreSpeech, MyST continued to out perform both.

4.1. Hidden Representations of Foundation Models

Foundation models, such as wav2vec, have demonstrated better performance in downstream tasks when finetuned, as compared to previous architectures before the adoption of transformers. However, a trend with advancements in these models has been to increase the size of models, implying the need for more parameters, and consequently, requiring more data to train these networks. Immediate issues such as the inability to scale, restriction of access to these models, and further opaqueness in our understanding will only be exacerbated by blindly following these recent trends. Therefore, we need to consider new directions that might point towards reducing model sizes, hopefully allowing us to understand how these models work.

In recent work by Pasad et al [9], they analyze representations produced by intermediate layers of pretrained wav2vec2 models. Through this analysis, they demonstrated that certain layers approximate are rich in information, with some approximating handcrafted features such as mel filterbanks. In subsequent research [10], Pasad shows that performance of a single best-performing layer can surpass using all layers in some downstream tasks, including ASR. In future work, this is a clear avenue to investigate in order to better tune speech models for tasks such as low-resource ASR.

5. Conclusion and Future Work

This project saw the implementation of ASR through Wav2Vec2 and its performance in Librispeech and MyST. Through augmentation, it is seen a limited dataset, MyST, can be further increased in order to improve WER. Afterwards, a language model such as a 5-gram can be incorporated to further improve the ASR. Overall, we saw an increase of 9.4% from the base model after all the improvements. For future work, increasing the amount of augmented data used for VLTP and pitch, or possibly changing the warping values could be tested to possibly see improvements in WER. On the neural network end, increasing the amount of iterations could be an easy way for seeing improvement in WER. More data was added in, but the same amount of iterations were kept so it is possible not all the added in augmented data was processed. There are many paths left to explore in order to improve the WER for MyST, with the most promising direction being the one mentioned in section 4.1.

6. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, 2019. [Online]. Available: http://arxiv.org/abs/1904.05862

[2] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[5] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, aug 2011. [Online]. Available: https://doi.org/10.1145/1998384.1998392

[6] E. Ma, "Nlp augmentation," https://github.com/makcedward/nlpaug, 2019.

[7] von Platen, Patrick, "Fine-Tune Wav2Vec2 for English ASR in Hugging Face with hugging-face Transformers," https://huggingface.co/blog/fine-tune-wav2vec2-english, 2023, [Accessed 13-Jun-2023].

[8] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, 2011, pp. 187–197. [Online]. Available: https://www.aclweb.org/anthology/W11-2123.pdf

[9] A. Pasad, J. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," *CoRR*, vol. abs/2107.04734, 2021. [Online]. Available: https://arxiv.org/abs/2107.04734

[10] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," 2023.