EXPLORING THE USAGE OF SPEECH FOUNDATION MODELS FOR LOW-RESOURCE ASR

Natarajan Balaji Shankar, Khushbu Pahwa, Eray Eren, Aurosweta Mahapatra

Department of Electrical and Computer Engineering, University of California Los Angeles

ABSTRACT

Self-supervised learning (SSL) has shown promising results in automatic speech recognition (ASR), particularly in the context of child speech recognition where labeled datasets are scarce. In this project we work on the recent research on SSL for child ASR and explore its impact on ASR system accuracy. Various SSL algorithms, such as contrastive learning and generative models, are used to train ASR systems on unlabelled data. The results of several studies demonstrate that SSL significantly improves the accuracy of child ASR systems. We use the Librispeech and MyST datasets for training and evaluating ASR models. Data augmentation techniques, including VTLP, SpecAugment, spectral warping, and pitch perturbation, are employed to enhance the performance of ASR models. The final part of the project includes a layerwise analysis of SSL models using Canonical Correlation Analysis to understand information propagation and feature extraction within the models. The findings provide valuable insights into the effectiveness and limitations of SSL for ASR tasks.

1. INTRODUCTION

1.1. Self-supervised learning

Self-supervised learning has emerged as a promising approach for automatic speech recognition (ASR) in recent years. This approach has shown significant success in improving the accuracy of ASR systems, particularly in the context of child speech recognition. In this document, we will review recent research on self-supervised learning for child ASR. This approach has shown significant success in improving the accuracy of ASR systems, particularly in the context of child speech recognition. One of the main challenges in child ASR is the lack of large-scale labeled datasets. This is where self-supervised learning comes in. It allows ASR systems to learn from unlabelled data, which is abundant in the case of child speech. Self-supervised learning algorithms use various techniques, such as contrastive learning, pretext tasks, and generative models, to learn from unlabelled data. Recent studies have shown that self-supervised learning can significantly improve the accuracy of child ASR systems. For instance, a study conducted by Liu et al [1] used a self-supervised learning approach to train an ASR system on child speech data. The results showed that the self-supervised learning approach outperformed the traditional supervised learning approach in terms of accuracy. Another study [2] used a contrastive learning approach to train an ASR system on child speech data. The results showed that the contrastive learning approach was able to improve the accuracy of the ASR system by a significant margin.

1.2. Speech Foundational Models for SSL

A foundation model, typically a large model trained on diverse data through self-supervision on a large scale, has garnered significant attention due to its remarkable enhancements in quality and emerging capabilities. In the field of speech, self-supervised pre-training of foundation models using vast amounts of unsupervised speech data has demonstrated impressive improvements in various speech recognition tasks. There are two primary approaches in self-supervised learning algorithms for speech. One approach involves directly reconstructing (APC, MPC) or predicting (Wav2vec) the input features. The other approach focuses on creating a BERT-style self-supervised learning model that bridges the gap between continuous speech signals and discrete text tokens, examples of which include Wav2vec 2.0, HuBERT, w2v-BERT, and BEST-RQ.

2. DATASET

In this study we experiment with Librispeech dataset and MyST dataset. The Librispeech dataset comprises read English speech from a diverse range of speakers.The dataset is split into different sections for training, development, and evaluation purposes. It provides several predefined splits, such as ”clean” and ”other” subsets. The ”clean” subset has been manually reviewed to ensure high-quality transcriptions, while the ”other” subset includes automatically generated transcriptions and may contain errors. LibriSpeech features contributions from approximately 2,500 different speakers, with a relatively equal gender distribution.Each utterance in the dataset is accompanied by a corresponding transcription, making it suitable for supervised learning approaches like ASR.

The MyST dataset contains spoken English from children aged between 6 and 14 years old. This dataset is aimed at the development of speech recognition models that are robust to variations in children's speech. The dataset is split into different sections for training, development, and evaluation purposes. It provides several predefined splits, such as "train-clean" and "dev-other" subsets. The "train-clean" subset has been manually reviewed to ensure high-quality transcriptions, while the "dev-other" subset includes automatically generated transcriptions and may contain errors.The MyST dataset features contributions from approximately 1,000 different child speakers, with a relatively equal gender distribution. Each utterance in the dataset is accompanied by a corresponding transcription, making it suitable for supervised learning approaches like ASR. It is commonly used by the research community to develop ASR models that are robust to variations in children's speech.

3. DATA AUGMENTATION

VTLP stands for Vocal Tract Length Perturbation. It is a technique used in speech processing to modify speech signals by adding perturbations to the vocal tract length. These perturbations help in enhancing the quality and intelligibility of the speech signals.VTLP augmentation is done by applying a perturbation filter to the speech signal. The filter is designed to mimic the effects of changes in vocal tract length that occur naturally in speech production. By adding these perturbations, the speech signal becomes more robust to noise and distortion.We use the NLPAug Python library to perform VTLP augmentation on speech signals.

SpecAugment (SpecAUg) is an augmentation technique commonly used in speech recognition tasks, particularly for acoustic modeling. SpecAugment operates on the spectrogram representations of audio data and introduces various modifications to enhance the training process and improve model generalization.SpecAugment provides several benefits in the context of acoustic modeling for speech recognition. It enhances model robustness to noise and variations, improves generalization, mitigates overfitting, augments data without additional recordings, enhances privacy and security, and reduces the reliance on manual transcription efforts. However, since SpecAUg does not introduce acoustic variability, its performance remains limited and this serves as a huge limitation. We apply SpecAug for all our experiments by setting the apply_spec_augment flag of the Wav2Vec2 model configuration as True.

Spectral warping [3] involves stretching or compressing the frequency axis of a spectrogram, which is a visual representation of the frequencies present in a signal. This can be achieved by applying a time-varying nonlinear transformation to the frequency axis of the spectrogram. The amount of stretching or compressing can be controlled by adjusting the parameters of the transformation. The spectral warping technique can be used to generate new training examples from existing ones by applying random transformations to the spectrograms. This results in a larger and more diverse training set, which can improve the generalization performance of the model.

Pitch Perturbation : We use pitch perturbation for adult speech to make it closer to child speech, effectively increasing the training data for child speech. We use SoX's "pitch" option for pitch perturbation, which shifts the original speech's pitch by "cents", i.e., 1/100th of a semitone. We experiment with different shift values, and find shifting adult speech's pitch up by 250 - 370 cents yields the best performance. For each utterance in the adult speech (Set A) dataset, we randomly pick a value between 250 - 370, and shift the utterance's pitch up by that value.

CycleGAN-VC2, is an improved version of CycleGAN-VC incorporating three new techniques: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN). We perform voice conversion between the adult and child speech using CycleGAN-VC2 and generated the converted voice samples. But owing to the limited compute, we were not able to finish the evaluation. This is aimed to convert large amount of adult speech corpus into children's speech so that the converted speech samples can be used for training children ASR.

4. ASR MODELLING PIPELINE

An ASR pipeline usually comprises multiple stages, which include acoustic modeling, language modeling (LM) decoding, and rescoring.

- Acoustic Modeling: The initial phase in the ASR pipeline involves acoustic modeling, where the audio input is processed to extract relevant characteristics like spectrograms or Mel-frequency cepstral coefficients (MFCCs). These features represent the acoustic properties of the speech signal.

- Language Modeling Decoding: Once the acoustic features are obtained, the ASR system performs decoding using a language model. The language model captures statistical patterns and probabilities of word sequences in a specific language. During decoding, the ASR system generates a set of potential word hypotheses that are likely to match the input speech. This process involves searching through a vast range of possible word sequences and determining the most probable sequence based on acoustic features and language model scores.

- Rescoring: Following the initial decoding step, the ASR pipeline often includes rescoring to refine the output and enhance accuracy. Rescoring involves reassessing the candidate word hypotheses using more advanced language models or other linguistic resources.

This step helps reduce errors and improve the final transcription.

5. EXPERIMENTS

5.1. Task1 : Fine-Tuning

For Task 1, we attempt to finetune wav2vec2 models of different sizes (base and large), and observe different patterns during the finetuning process. As we continue training we observe the the WER continues to lower till around 10k steps, a trend seen in Table 2. We primarily focus on optimizing the hyperparameters of learning rate and warmup steps. This can be seen for the base version of the model in Table 1. We also repeat the same experiments for the large model and compare the best performing base and large models in Table 3.

lr	dev-clean	dev-other	test-clean	test-other
5e-5	0.214	0.309	0.215	0.307
3e-5	0.222	0.31	0.224	0.313
7e-5	**0.199**	**0.288**	**0.202**	**0.291**

Table 1. Different learning rates for Librispeech base model

num_steps	dev-clean	dev-other	test-clean	test-other
900	0.301	0.399	0.305	0.401
10000	**0.199**	**0.288**	**0.202**	**0.291**

Table 2. Different no. of steps for Librispeech base model

Model_type	dev-clean	dev-other	test-clean	test-other
base 7e-5	0.199	0.288	0.202	0.291
large 3e-5	**0.142**	**0.2**	**0.143**	**0.199**

Table 3. Best base and large models on Librispeech

5.2. Task 2 : Data Augmentation

Similar to Task 1, we first focus on finetuning models on different hyperparameters, comparing the number of steps needed for convergence and the performance of the base and large models, which are seen in Tables 4, 5 and 6 respectively.

In addition to this, we also perform different data augmentation strategies to boost the WER of the model. The performance of models trained on different data augmentations is reported in Table 7.

5.3. Task 3 : Language Model Decoding & Rescoring

We first prepare different n-gram language models based on the provided 1hr librispeech text and use this for scoring potential hypotheses. We observed that the prepared 3-gram, 4-gram and 5-gram model were essentially identical and led to similar WER and thus for further experiments we report only the performance on the 4-gram model.

We then prepared a different 4-gram lm using the entirety of the librispeech train corpus, and observe thats its larger size led to a more efficient decoding on the best performing large model seen in Table 8.

Tables 9 and 10 concern themselves witht he effect of using out of domain data for deocding. Table 9 shows the effect of decoding on the best performing base model with out of domain data from wikipedia and riva, and observe that it does not appear to boost the wer. Table 10 uses a lm prepared from

lr	dev	test
1e-5	0.493	0.504
3e-5	0.403	0.41
5e-5	**0.379**	**0.381**
7e-5	0.384	0.391

Table 4. Different learning rates for MyST base model

num_steps	dev	test
1400	0.501	0.506
10000	**0.379**	**0.381**

Table 5. Different no. of steps for MyST base model

Model type	dev	test
base 5e-5	0.379	0.381
large 5e-5	**0.327**	**0.33**

Table 6. Best base and large models on MyST

data_aug	dev	test
no_aug	0.327	0.33
SP (0.4/1.6)	0.338	0.341
VTLP (200 epoch)	0.286	0.293
sp (0.9/1.1)	0.345	0.349
SFW (100 epoch)	0.296	0.306
all combined (50 epoch)	0.285	0.296

Table 7. Different Data Augmentation Techniques

LM type	dev-clean	dev-other	test-clean	test-other
no lm	0.142	0.2	0.143	0.199
libri_10h	0.125	0.178	0.126	0.177
libri_full	**0.119**	**0.173**	**0.119**	**0.171**

Table 8. Increasing Dataset size for n-gram LM

the MyST data and thus while there is some domain mismatch (child vs adult speech), it still benefits the decoding process more than the LMs prepared from textual domains.

LM type	dev-clean	dev-other	test-clean	test-other
no lm	0.199	0.288	0.202	0.291
libri_full	**0.164**	**0.244**	**0.167**	**0.248**
wikipedia	0.205	0.294	0.206	0.298
riva	0.204	0.292	0.205	0.297

Table 9. Out of Domain Datasets for n-gram LM

LM type	dev-clean	dev-other	test-clean	test-other
no lm	0.142	0.2	0.143	0.199
libri_full	**0.119**	**0.173**	**0.119**	**0.171**
myst_10h	0.135	0.188	0.134	0.187
myst_full	0.131	0.185	0.13	0.183

Table 10. Child vs Adult n-gram decoding of Librispeech

MLM type	dev-clean	dev-other	test-clean	test-other
no lm	0.142	0.2	0.143	0.199
bert	0.114	0.17	0.115	0.17
bert_ft	**0.106**	**0.161**	**0.107**	**0.161**

Table 11. MLM based rescoring of ASR

We also attempt rescoring based on Masked Language Models. Since masked language models do not have a simple loss to represent an entire sentence, we calculate the Pseudo log likelihood by sequentially masking each word in a candidate sequentially and summing up the loss from each of these sub candidates, and use the highest pll to select the hypothesis. We report the results in Table 11 of performing rescoring with Bert, and a version of Bert finetuned on libri_train.

LLM type	dev-clean	dev-other	test-clean	test-other
no lm	0.142	0.2	0.143	0.199
GPT2	0.112	0.167	0.111	0.164
GPT2_ft_10h	0.108	0.163	0.108	0.161
GPT2_ft_full	**0.102**	**0.158**	**0.103**	**0.156**

Table 12. Autoregressive LM based rescoring of ASR

Table 12 uses a rescoring approach where the top 16 most likely hypothesis from the n-gram lm model are fed back into a autogreressive large language mdoel, and the candidate with the lowest nll (negative log likelihood) is selected as the hypothesis. We attempt the experiment with 3 variants: the opensourced GPT2 model from Huggingfaces, a version of GPT2 finetuned on 1hr librispeech data and a version finetuned on the entire train section of librispeech.

LLM type	test-clean	test-other
no lm	0.143	0.199
Dolly	**0.106**	**0.16**

Table 13. Opensourced LLM based rescoring of ASR

Finally we also attempt to score hypothesis with a much larger language model Dolly, a distlled version of LLaMa finetuned with Alpaca dataset. We observe from Table 13 that its zero shot performance outperforms both GPT2 and BERT, but due to computational limitations we were unable to finetune this model, nor use it for decoding dev-clean and dev-other

LLM type	dev-clean	dev-other	test-clean	test-other
no lm	0.142	0.2	0.143	0.199
n-gram	0.119	0.173	0.119	0.171
LLM	0.112	0.167	0.111	0.164
LLM ft	**0.102**	**0.158**	**0.103**	**0.156**

Table 14. Comparison of different LM scoring techniques

The collated results from rescoring on Librispeech have been summarised in Table 14

We also repeat similar experiments from MyST, where the comparison of different n-gram models can be seen in Table 15, where we contrast using different subsets of Librispeech and MyST to decode the best performing large MyST model.

LM type	dev	test
no lm	0.327	0.33
libri_10h	0.301	0.306
libri_full	0.298	0.304
myst_10h	0.294	0.298
myst_full	**0.292**	**0.297**

Table 15. Child vs Adult ngram decoding of MyST

Table 16 contains the results of rescoring of the model on different finetuned versions of GPT2, and also contrasts the effect of finetuning GPT2 on Librispeech and using it for decoding MyST, as opposed to finetuning directly on MyST.

Finally, Table 17 highlights the effect of rescoring on boosting the performance of the best performing model that was trained with augmented data.

data_aug	dev	test
no lm	0.327	0.33
GPT2	0.299	0.302
GPT2_myst_10h	0.294	0.298
GPT2_myst_full	**0.289**	**0.294**
GPT2_libri_full	0.299	0.303

Table 16. Autoregressive LM based rescoring of MyST

Model Type	dev	test
no aug	0.327	0.33
with VTLP	0.286	0.293
with ngram	**0.257**	**0.264**

Table 17. Comparison of Data augmentation and LM rescoring

5.4. Task 4: Layerwise Analysis of SSL models

We conduct a layerwise analysis by extracting the representations from intermediate layers of the model and evaluating their quality on a downstream task of speech recognition.Canonical Correlation Analysis (CCA) is a statistical technique used to measure the linear relationship between two sets of variables. In the context of SSL, CCA is employed to assess the relationship between the representations learned by different layers of the SSL model and the labeled and unlabeled data. Layerwise analysis using CCA helps in understanding how information is propagated through the SSL model. It can reveal which layers capture discriminative features from labeled data and how subsequent layers exploit the un-labeled data to further refine the learned representations.

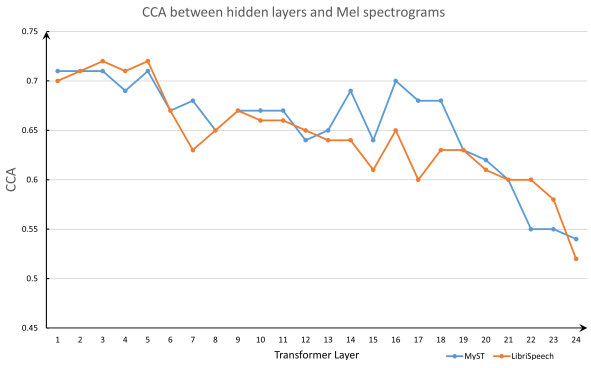

Fig. 1. CCA [4] between the each hidden transformer layers and mel-spectrograms

Figure 1 shows CCA [4] between the each hidden transformer layers and mel-spectrograms. To calculate CCAs, for each hidden transformer layer, randomly selected 50 development samples are concatenated in the temporal dimension, and their corresponding 80-dimensional mel-spectrograms are also concatenated similarly. For mel-spectrograms, librosa [5] library is used, and window length is set to 25 ms with 10 ms frame shift. We used the best performing models for extracting the hidden representations. For MyST dataset, the large wav2vec 2.0 model, finetuned with a learning rate of 0.00003 and 4200 iterations together with VTLP augmentation, is used. For LibriSpeech dataset, the large wav2vec 2.0 model, finetuned with a learning rate of 0.00003 and 10000 iterations, is used.

Figure 1 demonstrates that both MyST and LibriSpeech models are significantly correlated with acoustic features, mel-spectrograms. However, CCAs for different hidden layers are considerably different. Initial hidden layers (1-5) especially have high correlation with the acoustic features. On the other hand, CCAs drop significanly for early middle hidden layers (6-13). CCA scores tend increase again for the late middle layers (14-18). Then, CCAs drops dramatically for the final layers (19-24). Although MyST and LibriSpeech models have similar correlations with the acoustic features, MyST model has higher correlation for the late middle layers (14-18) compared to LibriSpeech model. We can argue that initial layers learn more acoustic-related features, and this relation decreases with the following layers though this decrease is non-linear.

6. CONCLUSION

In this project, we investigated the effectiveness of self-supervised learning (SSL) techniques in automatic speech recognition (ASR), specifically for child speech recognition. Our experiments showcased notable improvements in ASR accuracy through SSL methods. By fine-tuning the large Wav2Vec2 model with a learning rate of $3e^{-5}$ for 10K steps, we achieved a significant reduction in test-other word error rate (WER) from 0.313 to 0.199. Moreover, by combining SpecAug, SP, VTLP, SFW, and PP data augmentation techniques, we achieved a WER of 0.296 on the MyST test dataset. Additionally, leveraging the a finetuned Language Model for ASR rescoring yielded impressive WERs of 0.156 and 0.294 on Librispeech test-other and MyST test respectively. These findings demonstrate the effectiveness of SSL and data augmentation in improving child ASR accuracy.

7. ACKNOWLEDGEMENTS

We would like to thank Prof. Abeer Alwan for designing such an interesting project problem statement. We also like to thank the course TAs Vishwas and Ruchao for helping us in every stage of the project.

8. REFERENCES

[1] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma, "Self-supervised learning is more robust to dataset imbalance," *arXiv preprint arXiv:2110.05025*, 2021.

[2] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[3] John Makhoul and Lynn Cosell, "Lpcw: An lpc vocoder with linear predictive spectral warping," in *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1976, vol. 1, pp. 466–469.

[4] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," *CoRR*, vol. abs/2211.03929, 2022.

[5] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.