A Novel Noise-Robust Algorithm for Self-Supervised Pre-training of Speech Data

WEDNESDAY, APRIL 12, 2023

Article by Team AudioHeads, CS 753
{20d070009,20d070067,20d070047}

Speech recognition systems require a large amount of training data. Manual labelling of this data is a time-consuming and expensive endeavor, especially in the case of uncommon languages and specific professional domains. Additionally, a 10-fold increase in the training data is required to obtain a 40% reduction in the word error rate (WER) of a neural network model. Hence, it was necessary to develop methods of using unlabeled data to train speech recognition systems. Pretraining is one such method that learns speech representation from unlabeled data in a self-supervised way.

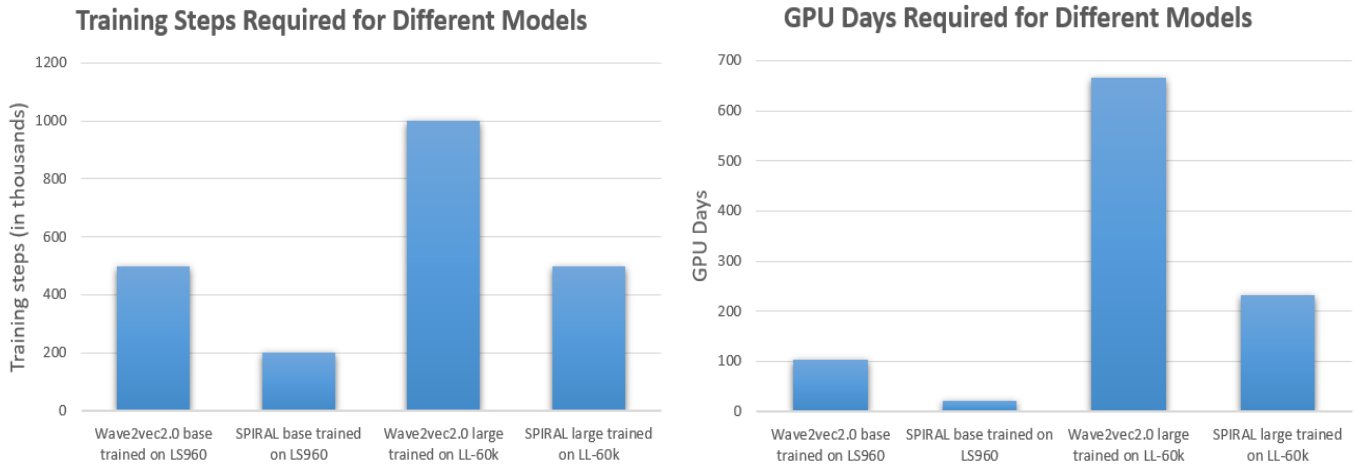
We introduce SPIRAL, a speech pretraining approach which works by learning denoising representation of perturbed data in a teacher-student framework. As compared to the state-of-the-art speech pre-training method wav2vec 2.0, SPIRAL requires less computation to train, making it a much better fit for modern machine learning hardware. Furthermore, it addresses the issue of noise-robustness, which is essential for real-world applications.

A Bit of History: The Self-Supervised Trend in Speech Recognition

Around 2015, researchers began to focus on techniques to utilize unlabeled data in speech recognition systems. Self-training involves initially training a teacher model with labeled data. The teacher model, usually combined with a language model, then produces pseudo-labels for unlabeled data. Labels and pseudo-labels are combined to train a student model, which is used as the teacher in the next iteration. A complementary method, pre-training uses self-supervision to learn speech representation from unlabeled data, followed by fine-tuning on labeled data.

The recently developed wave2vec2.0 model beat state-of-the-art results on the commonly used LibriSpeech (labeled) and LibriLight (unlabeled) datasets by using the Noisy Student Training method, which involves aggressive injection of noise into the student. To reduce the need for labeled data and to give reliable results on languages that were absent in pre-training, wave2vec2.0 uses powerful transformers and the masked language modeling method, in which some input tokens are masked, and the model must predict them. Training is done on quantized latent representations of data. Further research realized techniques like the offline clustering of similar frames to discover hidden units.

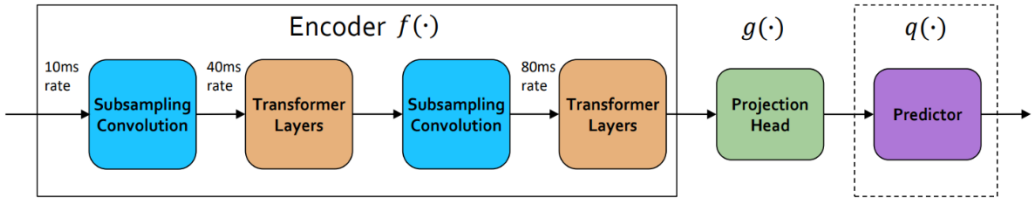
A big problem with previous models is the exorbitant cost of pre-training, which for the original wav2vec2.0 was around 16000 GPU-hours, making it hard to be used in low-latency automatic speech recognition systems. SPIRAL improves training time and allows end-to-end training with a single contrastive loss, without relying on discrete unit discovery techniques, such as the vector quantization method used in wave2vec2.0 or the iterative clustering process used in HuBERT.



Comparison of training steps and GPU days required by wave2vec and SPIRAL models. The base version of each model was trained on the LibriSpeech dataset while the large version of each model was trained on the LibriLight dataset

SPIRAL: Self-Supervised Perturbation-Invariant Representation Learning

Inspired by the ability of humans to communicate effectively in noisy environments, SPIRAL learns representation invariant to perturbation in a self-supervised way, to enhance speech applications using high-level representations. It uses a teacher-student framework to achieve this. The student network consists of a predictor, a projection head, and an encoder containing alternating subsampling convolutional and transformer layers. The teacher network has the same architecture except that it has no predictor.



The architecture of the student model in SPIRAL. The frame rate of input is denoted as '10/40/80 ms'. The dashed line indicates the optional predictor which can be removed with small performance degradation. The structure of the teacher model is the same but without the predictor. Image credit: Huang et. al.

Initially, a clean speech utterance is fed to the teacher network and its corresponding representation is obtained. The same utterance is then perturbed and fed to the student network. The student is trained to return a denoised representation of the perturbed utterance that resembles the teacher's representation of the clean utterance. Simultaneously, the teacher's weights are updated as a moving average of the student's weights over the past training steps.

An in-utterance contrastive loss is applied to prevent model collapse to trivial constant representation. Additionally, position randomization is applied by adding a random number of paddings on both sides of the teacher's input utterance to counter positional collapse, i.e., when the student exploits positional correlation in the teacher's representation to minimize loss, ignoring the actual content of the input utterance.

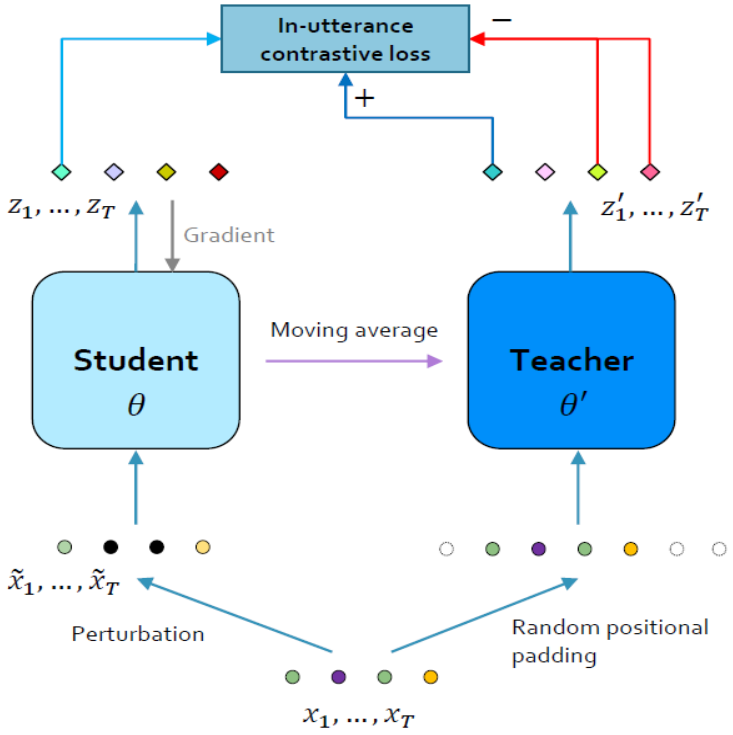


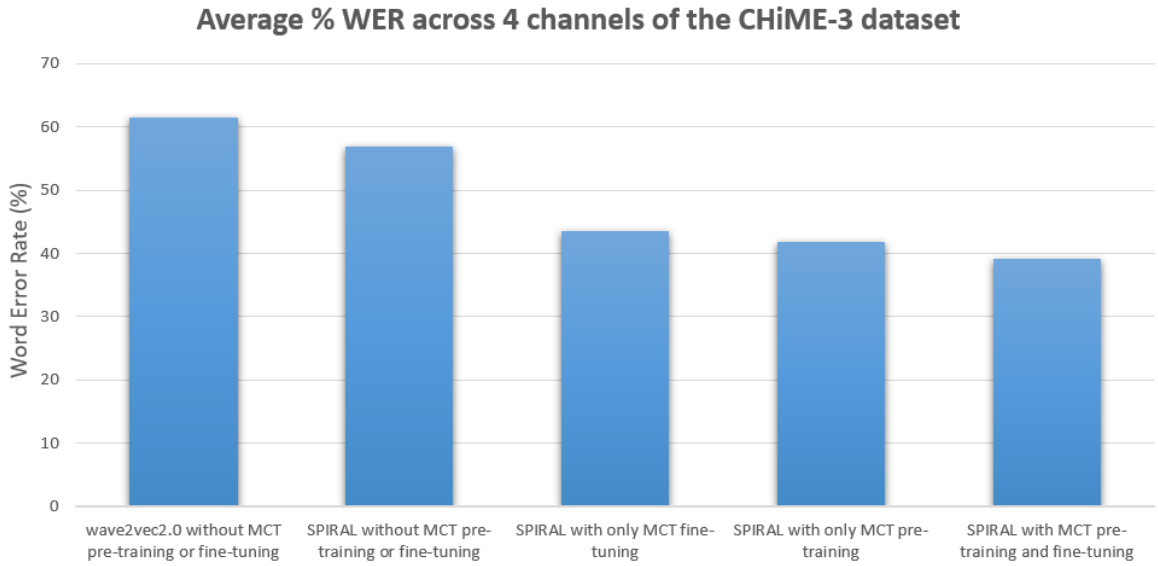
Illustration of SPIRAL architecture for speech pre-training. θ ' are the teacher weights which are updated as a moving average of the student weights θ . The input utterance is x_1, ..., x_T, the student's output representation is z_1, ..., z_T , and the teacher's output representation is z_1', ..., z_T'. Image credit: Huang et. al.

Finally, the SPIRAL model significantly reduces computation cost of pre-training with negligible performance degradation by using a gradual down-sampling strategy.

What's New: Noise-Robustness

Although most current speech recognizers give an acceptable recognition accuracy for clean speech data, there is a decline in their performance when they are used in real life situations, especially in noisy environments. For example, there is a drop in the performance of a conventional word recognizer which has been trained with clean speech data, from 100% accuracy to 30% accuracy in a car travelling 90 km per hour. In another study, the word error rate of a system trained under quiet conditions increases from 1% to over 50% in a cafeteria. Thus, environmental noise has become one of the major obstacles to the commercial use of speech recognition systems.

Multi-condition training is applied with the SPIRAL model to improve noise-robustness for downstream speech tasks. Multi-condition pre-training involves perturbing each input utterance to the student with several types of additive noise sampled from noise datasets. After multiple experiments on real (ChiME-3) and synthetic noisy test data, it was concluded that multi-condition pre-training gives more robustness to noise (Around 9.0%-13.3% relative word error rate reduction on real noisy test data) as compared to multi-condition training used solely in the fine-tuning stage.



Evaluation on noise-robustness of the models. Wav2vec2.0 BASE is used as baseline. The SPIRAL BASE models are pre-trained with LS-960 and fine-tuned with train-clean-100. The word error rate (WER) reported is the average of the WERs of microphone channels ch0, ch1, ch2, and ch5 of the CHiME-3 dataset.

In Conclusion: The Urgent Need for Computational Efficiency

In addition to the scarcity of labelled data, especially for less spoken languages, we can observe that as research progresses, speech processing models will become even larger, and only companies with large financial means will be able to afford to train them. Thus, despite incredible improvements in hardware in recent years, what we most need are improvements in the methods used to train these huge models, preferably to reduce their computational cost to a fraction of what they require today.

Acknowledgements

Research paper by Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu at Huawei Noah's Ark Lab. Article by Jakub Kaliski. Survey by Yifan Gong.