Project Proposal: Algorithms for speech and natural language processing

Project 1: *Self-supervised pretraining for phoneme recognition, and generalization on foreign languages*

Clément Apavou, Younes Belkada, Léo Tronchon, Arthur Zucker
MVA Master's program
École Normale Supérieure Paris-Saclay
firstname.lastname@ens-paris-saclay.fr

1. Proposal

Through this project, we will attempt to compare the results of our models on phone recognition tasks for different languages. Each one of us will be responsible for producing results for a global family of languages such as Indo-European, Sino-Tibetan, Afro-Asiatic, etc.

We will specifically define the languages that will be used for fine-tuning, therefore quantitatively and qualitatively compare the performances of multiple models – HuBERT, Wav2vec, and WavLM – on these languages.

We will explore how well self-supervised models are capable of generalizing to languages out of the training dataset.

2. Datasets

We will use the CommonVoices dataset[2] that is hosted on HuggingFace[1] in order to perform our studies. For each language, we will leverage the features from the pretrained models to train a small neural network with CTC. Then we will train it on 10 minutes, 1 hour and 10 hours of data and quantitatively compare the results using the Phoneme Error Rate on the provided testing set.

CommonVoices consists of a set of audio files together with the associated text scripts, in the Unicode format. Thus we have to convert the Unicode to the appropriate phonemes (which are in the IPA format and represent the phones in text). We will use the specifically designed *Phoible*[6] dataset for this task.

3. Languages

We will study the impact of choosing the right language for this task. Therefore, we will compare 4 different groups of languages sorted by their subjective *closeness* to English to see how much it correlates with model performance.

- European
- South American
- African
- Asian

For each group of languages, depending on the timeline and feasibility, we will select at least one language and run our experiments.

4. Distribution of Work

We will split the work among ourselves by languages and self-supervised pretraining methods, in order to get a similar amount of work.

5. Methods

We will compare Wav2Vec[7], HuBert[5] and WavLM[4]'s pretrained features trained on english datasets. Then we will compare how they perform on different languages. Ideally we would also like to try the very recent data2vec[3] method from Meta AI. Depending on the feasibility and timeline, we will chose at least one method that will be tested on each language.
We will then compare the results and observe which models generalize better on languages that differ from English.

References

[1] Mozilla common voice dataset on hugging face. Available at https://huggingface.co/datasets/common_voice. 1

[2] Mozilla common voice dataset, 2018. Available at https://commonvoice.mozilla.org/. 1

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. 1

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing, 2022. 1

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. 1

[6] Steven Moran and Daniel McCloy, editors. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena, 2019. 1

[7] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019. 1