école normale supérieure paris—saclay

MATHÉMATIQUES VISION APPRENTISSAGE

Self-supervised pretraining for phoneme recognition, and generalization on foreign languages

Algorithms for speech and natural language processing
Project Presentation

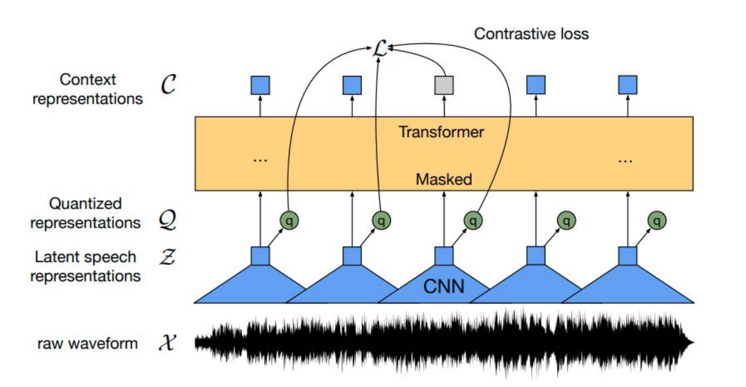
20/04/2022

APAVOU Clément
BELKADA Younes
TRONCHON Léo
ZUCKER Arthur

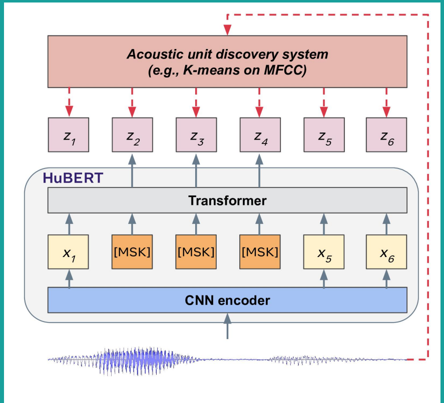
MVA Master's students
École Normale Supérieure Paris-Saclay
firstname.lastname@ens-paris-saclay.fr

Introduction

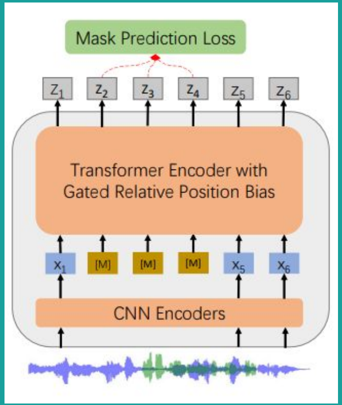
Recent advances in self-supervised learning for speech processing



Wav2Vec2



HuBERT

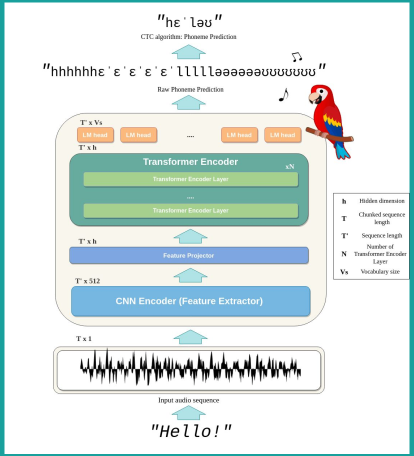


WavLM

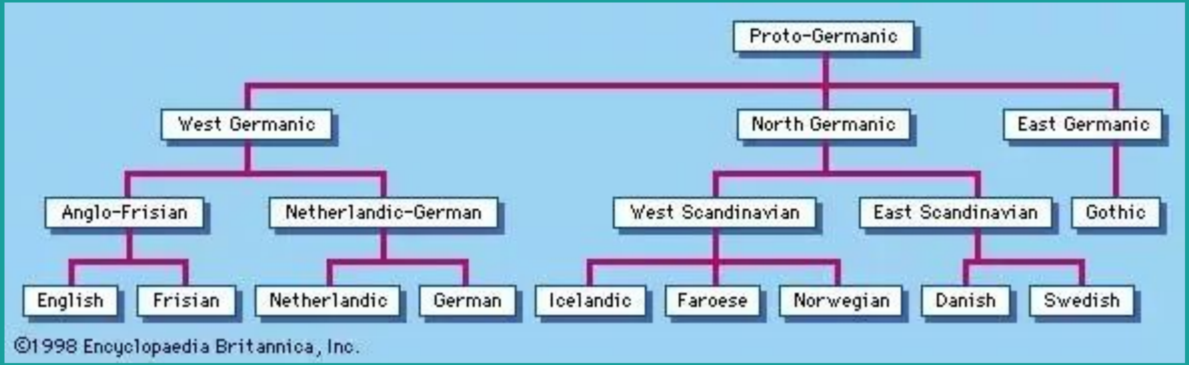
Can we use the learned features for phoneme recognition on various languages?

Goals / Problematics

What hypothesis would we like to confirm?



Our Phoneme Recognition pipeline with CTC

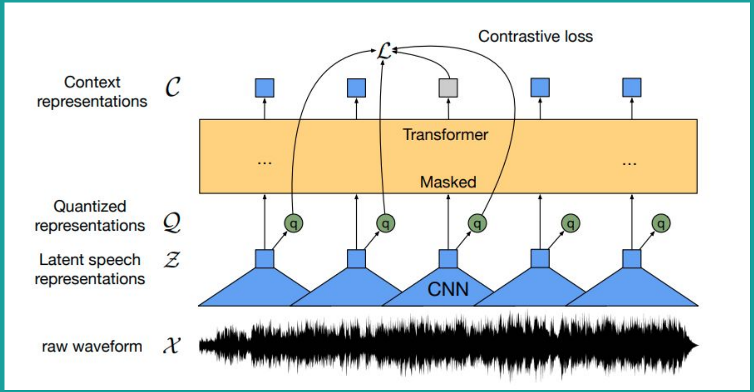


Which languages are close to English?

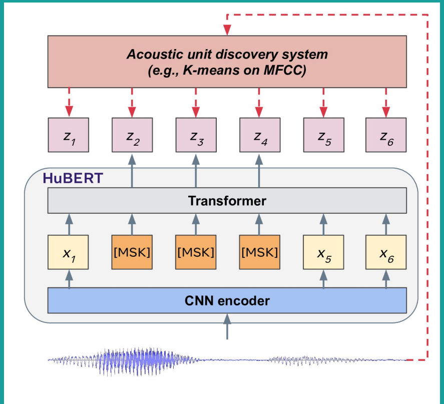
- How well can we perform phoneme recognition on other languages using pre-trained features on English?
- Is there a correlation between closeness to English and the model's performances?
- Which method allows to extract the best features for phoneme recognition?
- What is the influence of the abundance of training data on the performance of models?

Methods

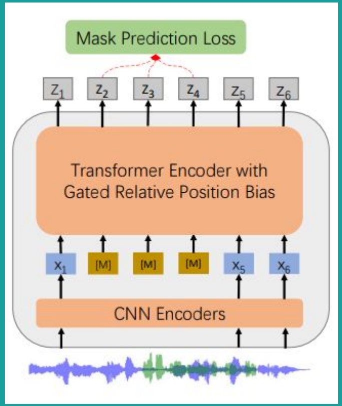
Pretrained models used



Wav2Vec2



HuBERT



WavLM

SoTA models on speech processing

1- Wav2vec2
2- HuBERT
3- WavLM

Pretrained models:

- Wav2vec2 Base, WavLM Base and HuBERT Large : 960 hours of Librispeech
- WavLM Large : MIX-94K (60K Libri-Light, 10K Gigaspeech and 24K VoxPopuli)

Models available on: 🤗

Datasets

Common Voice

Mozilla Common Voice is arguably one of the most famous open source dataset in ASR

Make use of Mozilla Common Voice on **5** languages:

1- Italian
2- Dutch
3- Swedish
4- Russian
5- Turkish

Language	number of phonemes
Swedish	40
Turkish	47
Russian	49
Dutch	52
Italian	60

Table 1. Number of phonemes of the 5 studied languages. We add 5 special tokens in addition to these phonemes: < s >, < /s >, < unk >, < pad > and |

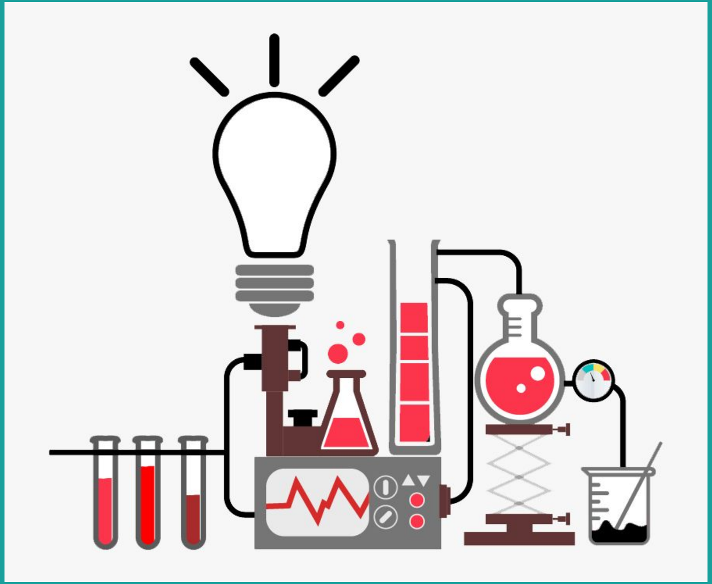
Dataset available on:

Converting transcripts to sequences of phonemes and tokenization:

- Phonemizer and espeak-ng backend
- Wav2Vec2PhonemeCTCTokenizer

Main Experiments

Fine-tune, frozen features and training data



Presenting our main experiments

+ Comparing *fine-tuning* HuBERT, WavLM and Wav2vec2 on our 5 languages
+ Comparing *the features learned by* HuBERT, WavLM and Wav2vec2 on our 5 languages (*i.e.,* freezing the network and training only the Linear head)
+ Comparing the impact of the amount of training data across the 3 models on Swedish

Fine-tuning

English closeness and training data

Main observations

1. Larger models are better
2. Closeness to English seems to be correlated to the PER of the models
   a. **Except** for Turkish
3. The amount of data seems to impact the performances as well

Language	Language Family	Proximity with English
Swedish	*North Germanic*	26.7
Dutch	*West Germanic*	27.2
Italian	*Romance*	47.8
Russian	*Est Slavic*	60.3
Turkish	*Turkic*	92.0

Table reporting closeness to English

Language	Training data (in hours)	Model	PER validation	PER test	Runs
Italian	62.34	Wav2Vec2 *Base*	19.05	17.95	Wav2Vec2_it
		HuBERT *Large*	**14.05**	**12.67**	Hubert_it
		WavLM *Base*	19.83	25.60	WavLM_it
Russian	15.55	Wav2Vec2 *Base*	32.16	31.66	Wav2Vec2_ru
		HuBERT *Large*	25.10	24.09	Hubert_ru
		WavLM *Base*	**20.25**	**18.88**	WavLM_ru
Dutch	12.78	Wav2Vec2 *Base*	16.18	20.83	Wav2Vec2_nl
		HuBERT *Large*	**12.77**	**16.49**	Hubert_nl
		WavLM *Base*	15.96	19.91	WavLM_nl
Swedish	3.22	Wav2Vec2 *Base*	26.50	24.16	Wav2Vec2_sv
		HuBERT *Large*	**21.77**	**19.38**	Hubert_sv
		WavLM *Base*	26.86	24.61	WavLM_sv
Turkish	2.52	Wav2Vec2 *Base*	19.62	19.03	Wav2Vec2_tr
		HuBERT *Large*	**15.51**	**14.19**	Hubert_tr
		WavLM *Base*	19.85	18.95	WavLM_tr
Average	-	Wav2Vec2 *Base*	22.70	22.73	-
		HuBERT *Large*	**17.84**	**17.36**	
		WavLM *Base*	20.55	21.59	

Table of experiments when models are **fine tuned**.

Frozen features

Comparison of pretrained methods

Language	Training data (in hours)	Model	PER validation	PER test	Runs
Italian	62.34	Wav2Vec2 *Base*	38.94	36.84	Wav2Vec2_it_tf_freezed
		WavLM *Base*	**27.29**	**25.98**	WavLM_it_tf_freezed
		HuBERT *Large*	23.85	21.15	Hubert_it_tf_freezed
		WavLM *Large*	**21.02**	**18.80**	WavLM_it_tf_freezed
Russian	15.55	Wav2Vec2 *Base*	50.11	48.69	Wav2Vec2_ru_tf_freezed
		WavLM *Base*	**40.66**	**38.76**	WavLM_ru_tf_freezed
		HuBERT *Large*	38.36	36.18	Hubert_ru_tf_freezed
		WavLM *Large*	**34.48**	**32.26**	WavLM_ru_tf_freezed
Dutch	12.78	Wav2Vec2 *Base*	40.15	39.23	Wav2Vec2_nl_tf_freezed
		WavLM *Base*	**34.94**	**35.67**	WavLM_nl_tf_freezed
		HuBERT *Large*	**27.62**	**26.68**	Hubert_nl_tf_freezed
		WavLM *Large*	27.71	27.19	WavLM_nl_tf_freezed
Swedish	3.22	Wav2Vec2 *Base*	50.30	45.23	Wav2Vec2_sv_tf_freezed
		WavLM *Base*	**43.65**	**40.55**	WavLM_sv_tf_freezed
		HuBERT *Large*	37.34	**32.68**	Hubert_sv_tf_freezed
		WavLM *Large*	**37.25**	33.14	WavLM_sv_tf_freezed
Turkish	2.52	Wav2Vec2 *Base*	53.92	52.08	Wav2Vec2_tr_tf_freezed
		WavLM *Base*	**47.18**	**45.53**	WavLM_tr_tf_freezed
		HuBERT *Large*	39.55	37.08	Hubert_tr_tf_freezed
		WavLM *Large*	**30.66**	**30.14**	WavLM_tr_tf_freezed
Average	-	Wav2Vec2 *Base*	46.68	44.41	-
		WavLM *Base*	**38.74**	**37.30**	
		HuBERT *Large*	33.34	30.75	
		WavLM *Large*	**30.22**	**28.31**	

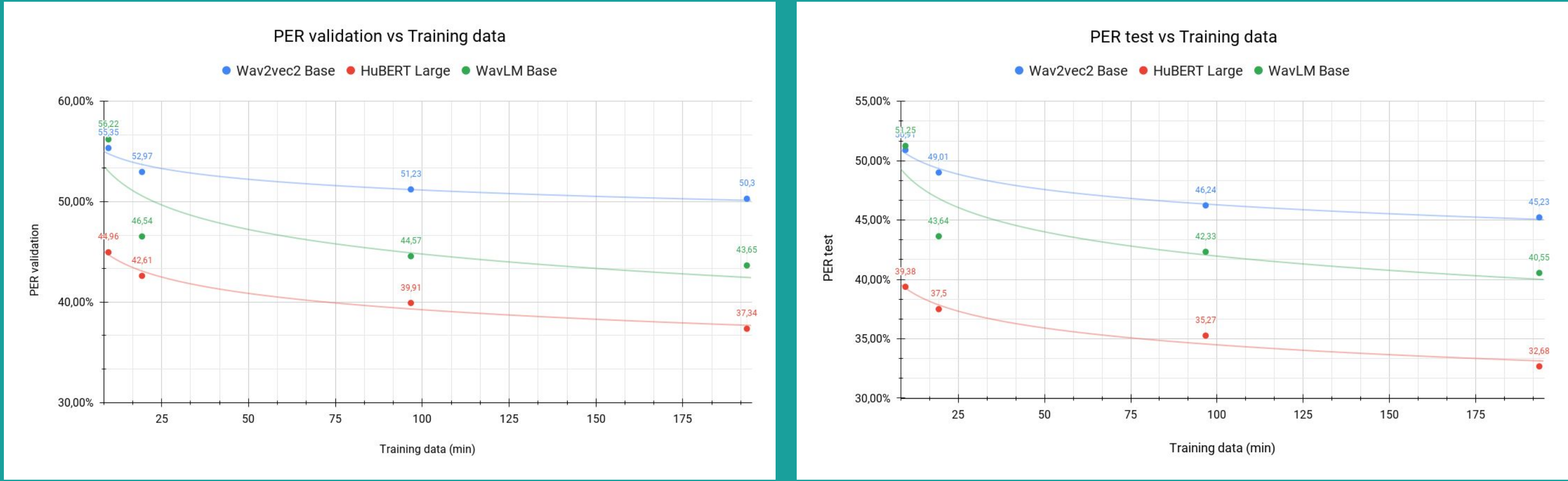
Table of experiments using **frozen features**.

Main observations

- PER higher than fine tune results
  - Best: 30.22% vs 17.84%
- Closeness to English definitely impacts the performance of the models
  - eg. Dutch > Russian > Turkish
- Wav2vec2 vs WavLM vs HuBERT
  - *Base:* WavLM > Wav2vec2
  - *Large:* WavLM > HuBERT
  - WavLM > Wav2vec2 and HuBERT

Training data

More is better but not so much



PER on the validation and test sets vs Training data for the Swedish language using frozen features.

amount of training data seems to be logarithmically correlated to the performance of the models
=> but not need a large amount of data to obtain decent results

Conclusions

Main conclusions

We have successfully built a framework for evaluating various pretrained models on phoneme recognition.

Main conclusions:

- **Closeness to English** impacts the performance of the model
- Overall **WavLM** seems to be **better than other pretrained methods**
- The **amount of training data** does not impact that much

Possible future works:

- What about other languages? *Japanese, Chinese, Hindi…*
- What about other new methods? *e.g. data2vec*

Code publicly available on github - Logs available on wandb

W&B

https://github.com/ASR-project/Multilingual-PR

https://wandb.ai/asr-project/test-asr?workspace=user-clementapa

THANKS FOR LISTENING !

Références

wav2vec 2.0 (2020) : https://arxiv.org/abs/2006.11477 , Baevski et al.
HuBERT (2021) : https://arxiv.org/abs/2106.07447 , Hsu et al.
WavLM (2022) : https://arxiv.org/abs/2110.13900 , Chen et al.