A study on fine-tuning wav2vec2.0 Model for the task of Mispronunciation Detection and Diagnosis

Linkai Peng[1], Kaiqi Fu[1], Binghuai Lin[2], Dengfeng Ke[1], Jinsong Zhang[1]

Beijing Language and Culture University,
Tencent Technology Co., Ltd, China

September 2021

Overview

1. **Introduction**
2. **Model**
3. **Experiments**
4. **Conclusion**

Introduction

End-to-end model
- avoid complicated modeling
- state-of-the-art performance
- require large amount of data

Mispronunciation Detection and Diagnosis (MDD)
- L2 data-scarce (annotations need the support of experts)

End-to-end MDD
- require large amount of data + data-scarce !
- add L1 data to train [Leung and Liu+ 19]
- add L1 data and pretrain [Yan and Wu+20][Yang and Fu+20]

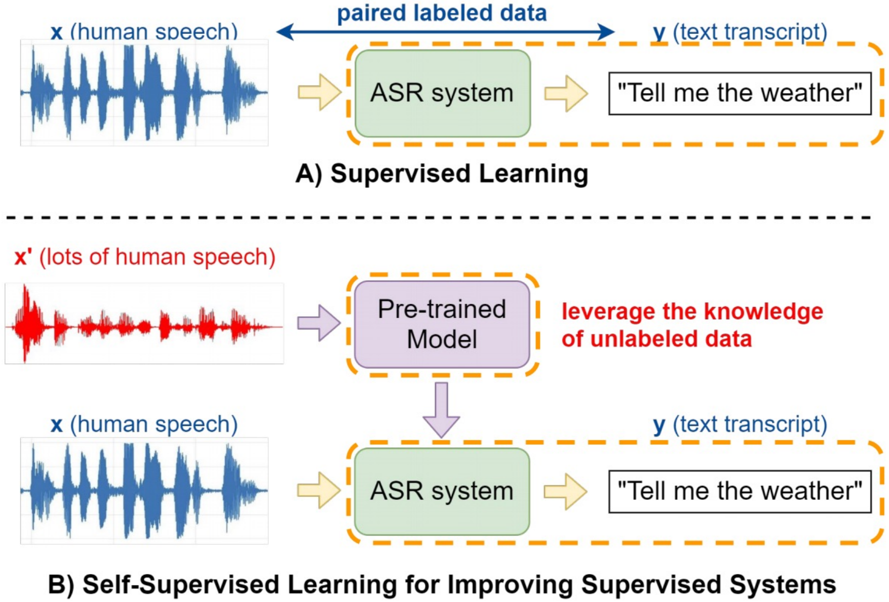
ASR:
- AISHEEL-1/2(178hrs/1000hrs)
- Common Voice(1400hrs)

MDD:
- L2-Arctic (3.66hrs annotated)

Introduction

Transfer learning - Self-Supervised Pretraining

> learn powerful context representation from unlabeled data



- **Autoregressive Predictive Coding (APC)**
  reconstruct the high-dimensional signal itself
  [Chung and Hsu+ 19]

- **Contrastive Predictive Coding (CPC)**
  distinguish the latent representation from a series of distractors
  [Baevski+][Liu and Yang+ 20]

* Figure come from https://andi611.github.io/files/20200514_asus_aics_SSL_for_speech.pdf

Introduction

Propose:

- **To introduce public pretraining model to MDD**
  wav2vec 2.0 (CPC-based)

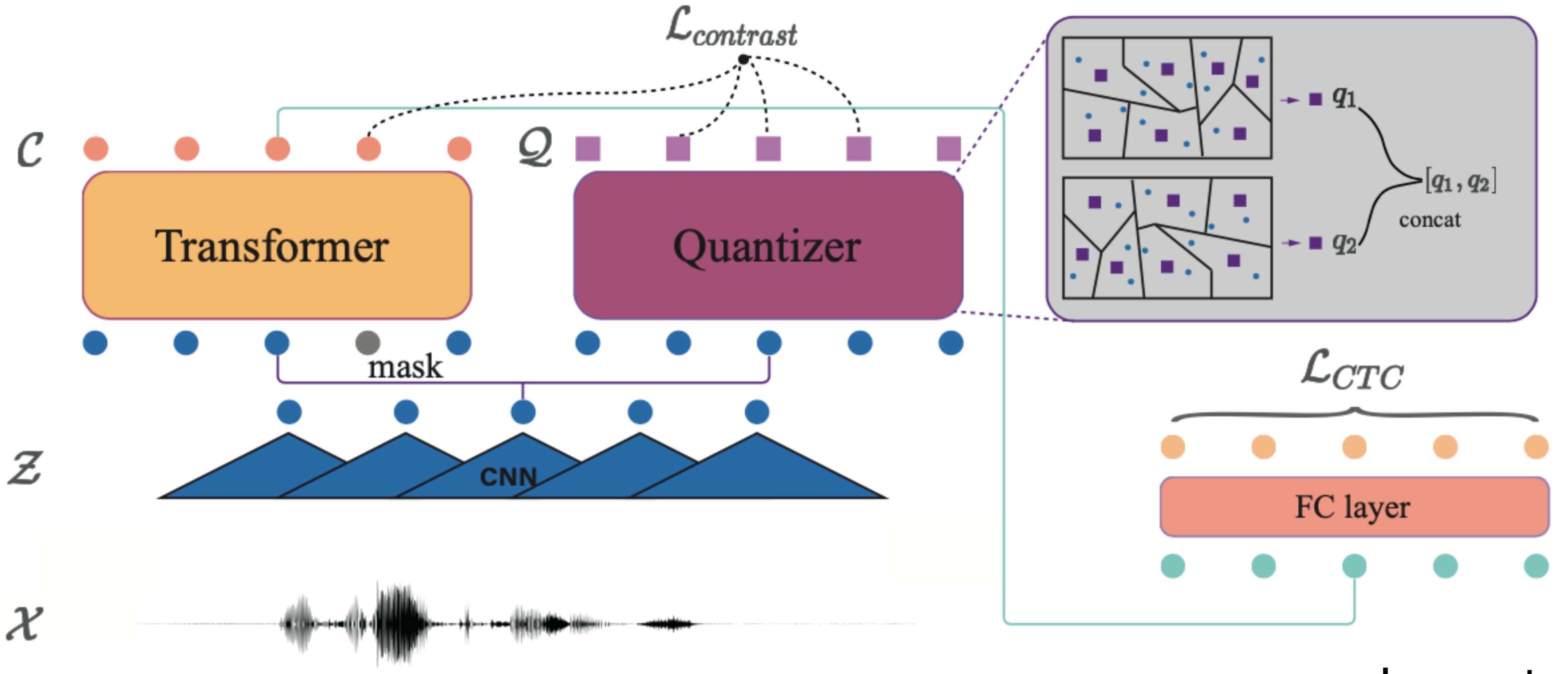
  > Recently has achieved state-of-the-art (SOTA) results in many tasks

- **To compare the performance under different conditions**
  monolinguistic/crosslinguistic models;
  low resource/extra low resource train data;

Overview

1. Introduction
2. **Model**
3. Experiments
4. Conclusion

Model



$$\text{L-constrast} = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

- **Pretrain**
  the objective is to distinguish the latent representation from other masked time steps. (L-constrast)
- **Fine-tuning**
  simply add a fully connected layer here to show the effectiveness of the SSP model on the MDD task. (L-CTC)

Overview

1. Introduction
2. Model
3. **Experiments**
4. Conclusion

Experiments - compare with other methods

- Data setup
  - Training data: TIMIT + L2-Arctic training set (following previous works)

Train	Dev	Tet
TIMIT train set 3.56 hours + L2-Arctic 2.5 hours	L2-Arctic 0.28 hours	L2-Arctic 0.88 hours

- Other configuration

  Wav2vec 2.0 XLSR (training with 50k hours data);
  Resample L2-Arctic data to 16kHz;
  Map the TIMIT 61-phone to 39-phone and then combine it into L2-Arctic phone set;
  Freeze wav2vec 2.0 for the first 10000 steps[total 18000];

Experiments - compare with other methods

Models	PR(%)	RE(%)	F1(%)
GOP[31]	35.42	52.88	42.42
CTC-ATT[7]	46.57	70.28	56.02
CNN-RNN-CTC+VC[32]	56.04	56.12	56.08
w2v2.0-XLSR	63.12	56.05	59.37
w2v2.0-XLSR(+TIMIT)	62.86	58.20	**60.44**

- XLSR achieves a 4.44% absolute improvement in F1 score (60.44% v.s 56.02%).
- Even without the use of the native speaker data, XLSR can still achieve a promising performance (59.37%)

Mispronunciation detection benefits from the general feature representation extracted from large amounts of unlabeled data

Experiments - comparison with different pre-trained models

Models	Data	Canonicals		Mispronunciations			F1	PER
		True Accept	False Rejection	False Accept	True Rejection			
					Corroct Diag.	Diag. Error		
w2v2.0-LARGE	-	94.12% (24226)	5.88% (1514)	49.53% (2113)	65.86% (1418)	34.14% (735)	54.28%	16.97%
w2v2.0-LV60	-	94.01% (24198)	5.99% (1542)	43.37% (1850)	68.08% (1645)	31.91% (771)	58.75%	16.01%
w2v2.0-XLSR	-	94.57% (24343)	5.43% (1397)	43.95% (1875)	65.75% (1572)	34.25% (819)	**59.37%**	**15.43%**

Note:
- LARGE: 960h hours
- LV60: 53,200+ hours
- XLSR: 53 languages, 56000 hours

- LARGE v.s LV60
  Mispronunciation detection benefits from the general feature representation extracted from large amounts of unlabeled data

- LARGE v.s XLSR
  BG: language learners will transfer the phonetic phenomenon o their mother tongue to second language learning
  the multilingual pre-trained model can transfer cross-language information for pronunciation evaluation*

* Note: the impact of multilingual training, unrelated languages and simply training on more data should be disentangled.

Experiments – training on few data

Models	Data	Canonicals		Mispronunciations			F1	PER
		True Accept	False Rejection	False Accept	True Rejection			
					Corroct Diag.	Diag. Error		
w2v2.0-LARGE	-	94.12% (24226)	5.88% (1514)	49.53% (2113)	65.86% (1418)	34.14% (735)	54.28%	16.97%
w2v2.0-LV60	-	94.01% (24198)	5.99% (1542)	43.37% (1850)	68.08% (1645)	31.91% (771)	58.75%	16.01%
w2v2.0-XLSR	-	94.57% (24343)	5.43% (1397)	43.95% (1875)	65.75% (1572)	34.25% (819)	**59.37%**	**15.43%**
w2v2.0-XLSR	-33%	94.11% (24156)	5.89% (1584)	41.23% (1802)	69.13% (1712)	30.87% (752)	59.27%	-
w2v2.0-XLSR	-66%	93.35% (23048)	6.65% (2692)	46.06% (1592)	64.67% (1870)	35.33% (804)	55.52%	-
w2v2.0-XLSR	+TIMIT	94.30% (24273)	5.70% (1467)	41.80% (1783)	70.72% (1756)	29.28% (727)	60.44%	16.20%

Data

	Train	Dev	Test
Default	2.50	0.28	0.88
-33%	1.49	0.37	0.88
-66%	0.73	0.19	0.88
+TIMIT	6.07	0.28	0.88

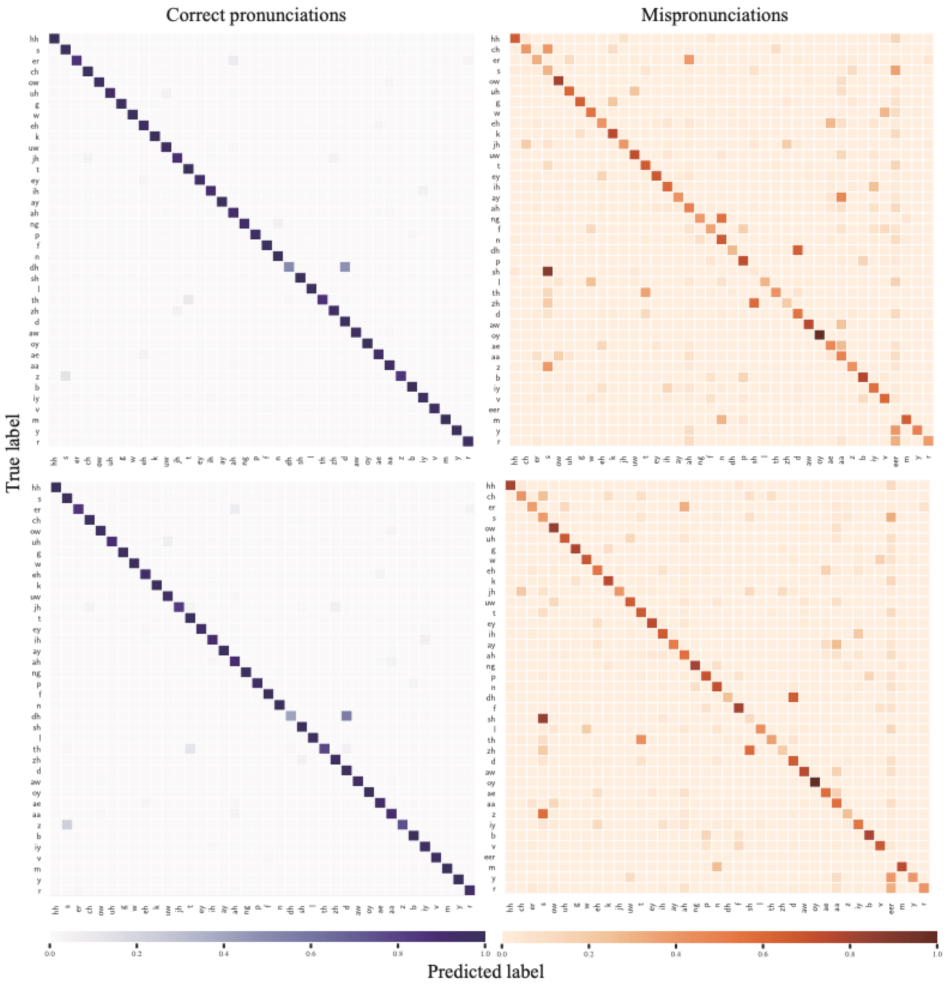
# of speakers for each language on the training dropped from

-33%: 3 -> 2

-66%: 3 -> 1

The feature representations generated from wav2vwc2.0 can rapidly generalize on MDD tasks even when access to annotated data is limited.

Experiments – training on few data



default (up) and -66% (down)
The diagonal cells indicate

- True Accept for the correct pronunciations
- False Accept for the mispronunciations

The model using less annotated data can retain most of the ability to distinguish phones

Conclusion

- The self-supervised pre-training model can take advantage of unlabeled data and provide useful speech representations for the MDD task.

- The feasibility of ultra-low resource MDD

Future work

- **provide specific diagnostic information**
- **children’s speech assessment.**

Thank you for your attention

Any questions?

Reference

- [Leung and Liu+ 19] W.-K. Leung, X. Liu, and H. Meng
Cnn-rnn-ctc based end-to- end mispronunciation detection and diagnosis,
ICASSP
- [Yan and Wu+20] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen
An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling
in Proc. Interspeech 2020,
- [Yang and Fu+20] L. Yang, K. Fu, J. Zhang, and T. Shinozaki
Pronunciation erroneous tendency detection with language adversarial represent learning
Proc. Interspeech 2020
- [Chung and Hsu+ 19] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. R. Glass
An unsupervised autoregressive model for speech representation learning
in *INTERSPEECH*, 2019.
- [Baevski+] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli
wav2vec 2.0: A framework for self-supervised learning of speech repre- sentations,
Advances in Neural Information Processing Systems, vol. 33, 2020.
- [Liu and Yang+ 20]
Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders
in *ICASSP 2020-2020*