Training wav2vec on Multiple Languages From Scratch

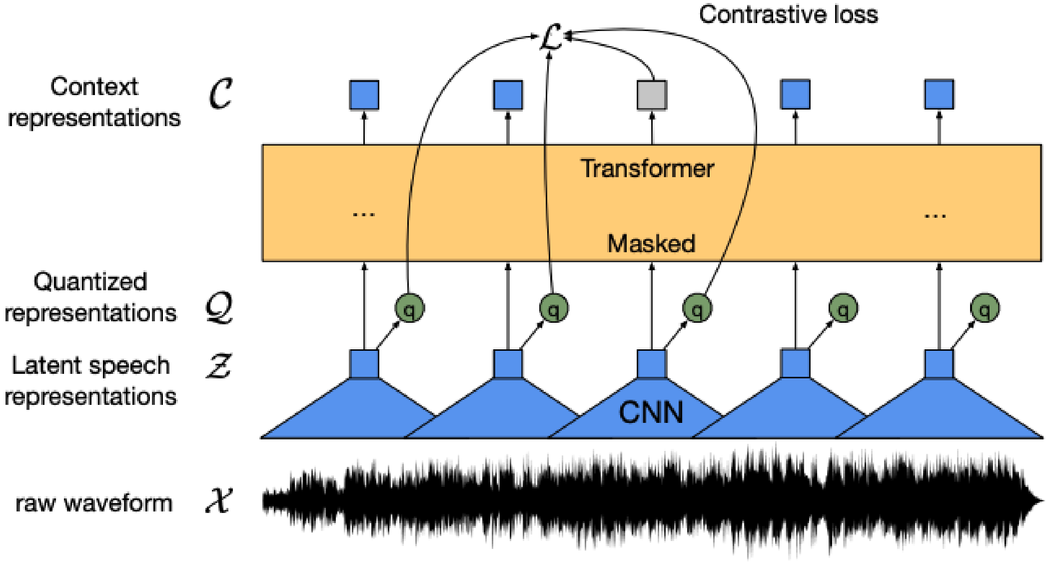
Heting Gao, Mahir Morshed, Shuju Shi, Liming Wang, Junkai Wu

Introduction

- Large amount of parallel speech-text data not available in most languages
- wav2vec: a new paradigm of training an ASR system by splitting the training process into two stages:
  - Self-supervised pretraining (only unlabeled audio is required)
  - Low-resource finetuning (small amount of parallel speech-text data is required)

Introduction

- Model Architecture
  - Multi-layer convolutional feature encoder
  - Quantization module to discretize the features into codewords in a codebook
  - Transformer encoder to output contextual representation of each frame
- Pretraining to predict the codeword of the current frame using content representation
  - Contrastive loss with negative sampling
  - Codebook diversity loss
- Finetuning
  - Additional linear layers and CTC loss
  - Freeze the pretrained wav2vec and finetune only the linear layers



Introduction

- Previous works:
  - wav2vec 2.0[1] only trained the model on English
  - XLSR-53[2] trained a multilingual wav2vec on a 53-language dataset
    - Multilingual wav2vec has a better cross-lingual performance
  - VoxPopuli[3] released a large-scale multilingual speech corpus
    - 23 languages of the European Union
      - of which 16 partially have transcriptions
    - Pretrained wav2vec models provided
- Objectives:
  - How does wav2vec work with other languages (such as Asian languages)?
  - How does a wav2vec pretrained on different languages affect performance?

[1] doi:10.48550/arXiv.2006.11477 [2] doi:10.48550/arXiv.2006.13979 [3] doi:10.18653/v1/2021.acl-long.80

Datasets considered
(chosen languages highlighted)

- Multilingual LibriVox
  - **English (en)** (36000 h), German (1691 h), Dutch (1264 h), French (897 h), Spanish (735 h), Italian (220 h), Portuguese (136 h), Polish (91 h)
- LibriLight for **English (ft-en)** (10h used)
- LaboroTVSpeech for **Japanese (ja)** (2000 h)
- Corpus of Spontaneous **Japanese (ft-ja)** (661 h)
- Babel (200 h per language)
  - Bengali, Vietnamese, Zulu, Amharic, Javanese, Georgian, Cantonese, Lao
- GlobalPhone (20 h per language)
  - Czech, French, **Mandarin (ft-zh)**, Thai, German, Portuguese, Turkish, **Bulgarian (ft-bg)**, Croatian, Spanish, Polish
- OpenSLR (all but the last two from Google)
  - Javanese (295 h), Sundanese (332 h), Sinhala (225), Bengali (215 h), Nepali (154 h), Korean (51 h), Kazakh (332 h)
- Europarl-ST
  - English (637 h), French (176 h), German (153 h), Italian (181 h), Spanish (116 h), Portuguese (82 h), Polish (151 h), Romanian (108 h), Dutch (38 h)
- United Nations Proceedings Speech (~1000h per language)
  - English, **Mandarin (zh)**, **Standard Arabic (ar)**, French, **Russian (ru)**, Spanish
- Common Voice
  - Kinyarwanda (2000 h), Esperanto (1300 h), Catalan (1200 h), Belarusian (1000 h)
- VoxPopuli (2.7k to 24.1k h)
  - English, German, French, Spanish, Polish, Italian, Romanian, Hungarian, Czech, Dutch, Finnish, Croatian, Slovak, Slovene, Estonian, Lithuanian, Portuguese, **Bulgarian (bg)**, Greek, Latvian, Maltese, Swedish, Danish
- GALE Broadcast News Datasets for **Standard Arabic (ft-ar)** (~120 h for Phase 3 Part 2)
- **Russian (ft-ru)** LibriSpeech (~100 h)

Experiment (English Baseline)

- Hardware
  - NCSA's HAL cluster
    - 4 x 16 GB NVIDIA V100 GPU
- wav2vec model settings
  - wav2vec base model has max token per batch: 1M (originally was 1.4M)
  - XLSR-53 model has max token per batch: 600k (originally was 1.28M)
  - Update frequency is 16 to simulate 64 GPU training (2 weeks)
- English baseline
  - Better performance (unit error rate) when finetuned on English
  - Slightly worse performance when finetuned on Bulgarian

	Validation UER	Validation WER	Test UER
official-en-ft-en	-	10.9	-
en-ft-en	3.51	**9.87**	2.97
official-en-ft-bg	3.34	17.37	**3.31**
en-ft-bg	3.48	17.68	3.47

Experiment (Mono- vs Cross- vs Multi-lingual)

- Three settings
  - Monolingual finetuning
  - Cross-lingual finetuning (English)
  - Multilingual finetuning (XLSR-53)
- Monolingual finetuning > Multilingual finetuning > Cross-lingual finetuning
  - English and Russian excepted

Mono	Train Loss	Valid UER	Valid WER	Test UER	Cross	Valid UER	Valid WER	Test UER	Multi	Valid UER	Valid WER	Test UER
en-ft-en	0.07	3.51	9.87	2.97	en-ft-en	3.51	9.87	2.97	xlsr-ft-en	1.91	6.58	**1.91**
bg-ft-bg	1.53	1.85	8.78	**1.89**	en-ft-bg	3.48	17.68	3.47	xlsr-ft-bg	2.67	13.79	2.70
zh-ft-zh	2.00	10.43	-	**10.60**	en-ft-zh	15.20	-	15.41	xlsr-ft-zh	14.15	-	14.56
ru-ft-ru	1.95	5.57	23.06	6.98	en-ft-ru	5.57	27.92	5.59	xlsr-ft-ru	5.84	28.21	**4.84**
ar-ft-ar	1.77	3.62	12.32	**3.45**	en-ft-ar	6.49	20.41	5.47	xlsr-ft-ar	4.67	17.56	4.58
jp-ft-jp	1.98	10.38	-	**9.91**	en-ft-jp	16.08	-	16.33	xlsr-ft-jp	14.67	-	14.35

Experiment (Grapheme vs IPA)

- Convert graphemes into International Phonetic Alphabet (IPA) using LanguageNet Grapheme-to-Phoneme Transducers (g2ps)
- Expected IPA transcripts to have lower error rates
  - Turns out not to be the case generally
    - Bulgarian and Mandarin have much lower error rates using graphemes
    - English, Russian and Arabic have slightly lower error rates using graphemes
    - Japanese has much lower error rate using IPA

Mono	Test UER	Test UER IPA	Cross	Test UER	Test UER IPA	Multi	Test UER	Test UER IPA
en-ft-en	**2.97**	3.19	en-ft-en	**2.97**	3.20	xlsr-ft-en	**1.91**	2.29
bg-ft-bg	**1.89**	11.28	en-ft-bg	**3.47**	14.86	xlsr-ft-bg	**2.70**	17.38
zh-ft-zh	**10.60**	15.03	en-ft-zh	**15.41**	18.05	xlsr-ft-zh	**14.56**	16.10
ru-ft-ru	**6.98**	7.87	en-ft-ru	**5.59**	6.10	xlsr-ft-ru	**4.84**	5.26
ar-ft-ar	**3.45**	4.20	en-ft-ar	**5.47**	6.55	xlsr-ft-ar	**4.58**	5.24
jp-ft-jp	9.91	**3.32**	en-ft-jp	16.33	**4.98**	xlsr-ft-jp	14.35	**4.51**

Experiment (Mandarin)

- Extra experiments on Mandarin Chinese
  - Chinese characters
  - IPA with or without tone
  - Pinyin with or without tone
- Converting to Pinyin phonemes can greatly reduce the error rate
- g2ps probably contain errors when transducing from Chinese characters to IPA
- wav2vec can capture tone information very well

zh-ft-zh	Test UER
char	10.60
IPA	15.03
IPA w/o tone	14.56
Pinyin	2.96
Pinyin w/o tone	**2.31**

en-ft-zh	Test UER
char	15.41
IPA	18.05
IPA w/o tone	16.94
Pinyin	3.80
Pinyin w/o tone	**2.97**

xlsr-ft-zh	Test UER
char	14.56
IPA	16.10
IPA w/o tone	15.44
Pinyin	4.02
Pinyin w/o tone	**2.79**

Experiment (Japanese Kana)

- Extra experiments on Japanese Kana
- Use wav2vec pretrained on Japanese, English, Mandarin Chinese, Spanish and XLSR-53
- Finetune on Japanese
  - Cross-lingually, English > Spanish > Mandarin Chinese
  - Multilingual XLSR-53 is better than the cross-lingual models.

Kana	Test UER
jp-ft-jp	*5.08*
en-ft-jp	8.18
es-ft-jp	8.60
zh-ft-jp	9.66
xlsr-ft-jp	**7.16**