Building a multilingual Speech-to-IPA system

[bɪldɪŋ ə mʌɫtɪlɪŋgwəɫ spit͡ʃ tʰə foʊnim sɪstəm]

Members: Chihiro Taguchi, Yusuke Sakai, Lusine Vanyan, Aditi Swarnkar

Mentor: Parisa Haghani

Instructions

- Please **edit directly on this google slide deck**. During the presentation, you will use a provided laptop for the presentation.

- The final presentation should consist of **3 min presentation + 1-2min QA from judges**. Please stick to the time as we will stop presentations that exceed 5 min.

- In your presentation please consider the following:

- Goal of the project and what social or economic impact could it create
- What it makes interesting and/or innovative ?
- Challenges you have overcome
- What have you learned from it ?
- What makes the project special or gives your proposal an edge over similar solutions in the market ?

TIPS and guidelines

- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

Introduction: Why Speech-to-IPA?

Problem:

- Transcription is **time-consuming** in language documentation
- ASR for IPA is **understudied** and **underdeveloped**

Solution:

- **Build a speech-to-IPA model** for any languages

Social Impact:

- **Efficient documentation** of endangered languages

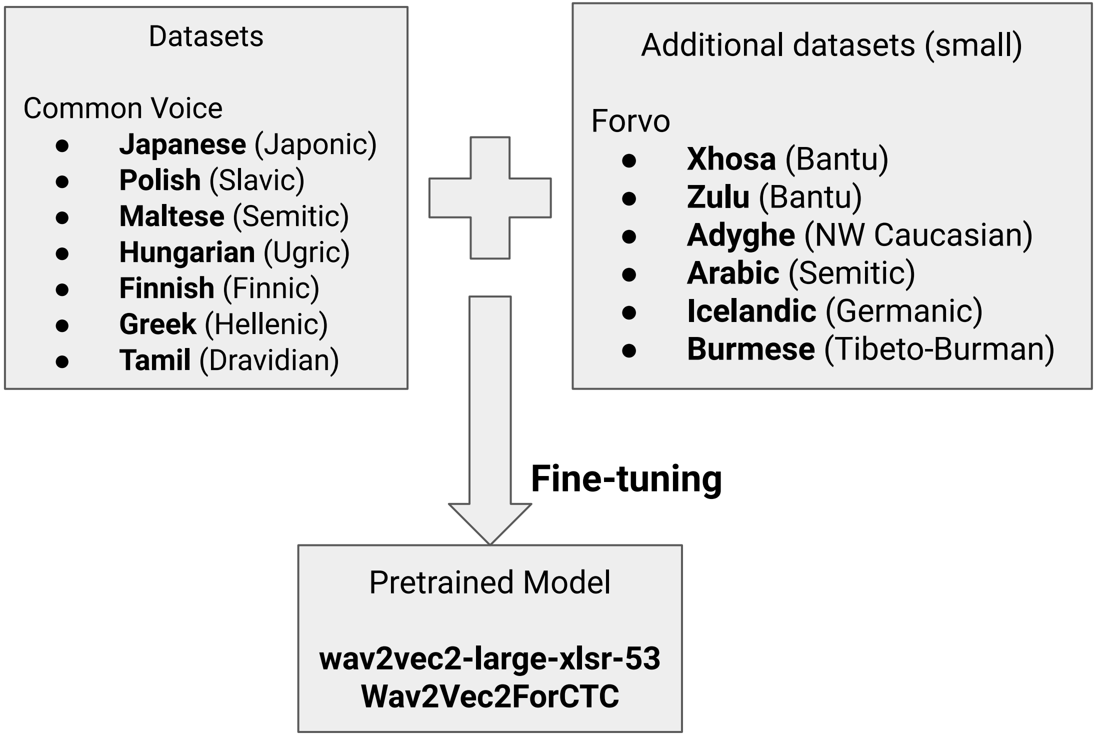
Method

- Pre-trained model
  - **wav2vec2-large-xlsr-53** by Facebook
- Fine-tuning:
  - **CTC** (Connectionist Temporal Classification)
- Datasets
  - **Common Voice** (Japanese, Polish, Maltese, Hungarian, Finnish, Modern Greek)
  - **Forvo** (Xhosa, Zulu, Adyghe)
- Evaluation:
  - Character Error Rate (CER) or our **new metrics**

Low-resource problem

- Few high-quality speech-to-IPA data
- Workaround
    - **Orthography-to-IPA** (Common Voice)
        - Off-the-shelf modules: not very accurate
        - + manually prepared rules (only “spelled-as-pronounced” langs)
    - **Create dataset manually**
        - Audio: Forvo
        - Manually annotate phonetic transcription

Setup (Goal)



Datasets

Common Voice

- **Japanese** (Japonic)
- **Polish** (Slavic)
- **Maltese** (Semitic)
- **Hungarian** (Ugric)
- **Finnish** (Finnic)
- **Greek** (Hellenic)
- **Tamil** (Dravidian)

Additional datasets (small)

Forvo

- **Xhosa** (Bantu)
- **Zulu** (Bantu)
- **Adyghe** (NW Caucasian)
- **Arabic** (Semitic)
- **Icelandic** (Germanic)
- **Burmese** (Tibeto-Burman)

Fine-tuning

Pretrained Model

wav2vec2-large-xlsr-53
Wav2Vec2ForCTC

Baevski et al. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. https://arxiv.org/abs/2006.11477

3000 km
2000 mi
Leaflet | © OpenStreetMap contributors

IPA coverage: consonants (pulmonic)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Nasal	m̥ m	ɱ		n̥ n		ɳ̊ ɳ	ɲ̊ ɲ	ŋ̊ ŋ	ɴ		
Plosive			t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ	ʡ*	ʔ
Sibilant affricate					t͡s d͡z	t͡ʃ d͡ʒ	t͡ɕ d͡ʑ				
Non-sibilant affricate	p͡ɸ* b͡□*	p̪͡f* b̪͡v*	t͡θ* d͡ð*	t͡ɹ̝̊* d͡ɹ̝*	t̠͡ɹ̠̊˔* d̠͡ɹ̠˔*	c͡ç^ ɟ͡ʝ^	k͡x^ ɡ͡ɣ^		q͡□^ ɢ͡ʁ*	ʡ͡ʜ* ʡ͡ʢ*	ʔ͡h*
Sibilant fricative					s z	ʃ ʒ	ʂ ʐ	ɕ ʑ			
Non-sibilant fricative	ɸ β	f v	θ ð	θ_ð̠	ɹ̠̊˔* ɹ̠˔*	ɻ̊˔* ɻ˔*	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ			ʔ̞*
Tap/flap	ⱱ̟*	ⱱ*		ɾ̥ ɾ		ɽ̊* ɽ			ɢ̆^	ʡ̆*	
Trill	ʙ̥* ʙ+			r̥ r	ɽ̊r̥* ɽr*				ʀ̥* ʀ	ʜ+ ʢ+	
Lateral affricate				t͡ɬ d͡ɮ			t□* ɖɭ˔*	cʎ̥˔* ɟʎ̝*	kʟ̝̊* ɡʟ̝*		
Lateral fricative				ɬ ɮ		□* ɭ˔*	ʎ̝̊* ʎ̝*	ʟ̝̊* ʟ̝*			
Lateral approximant				l		ɭ	ʎ	ʟ*	ʟ̠*		
Lateral tap/flap				ɺ̥* ɺ^		ɭ̥̆* ɭ̆^	ʎ̆*	ʟ̆*			

IPA coverage: consonants (others)

Non-pulmonic		bilabial	labio-dental	dental	alveolar	post-alveolar	retroflex	palatal	velar	uvular	epiglottal
ejective	stop	**p’**			**t’**		ʈ’*	c’*	**k’**	q’	ʡ’*
	affricate			t̪θ’*	**t͡s’**	**t͡ʃ’**	**ʈ͡ʂ’**		kx’*	q͡χ’*	
	fricative	ɸ’*	f’*	θ’*	s’	**ʃ’**	**ʂ’***	**ɕ’**	x’*	χ’*	
	lateral affricate				**tɬ’**						
	lateral fricative				**ɬ’**						
Click	tenuis	ʘ*		**ǀ**	**ǃ**			ǂ			
	voiced	ɡʘ*		ɡǀ	**ɡǃ**			ɡǂ			
	nasal	ŋʘ*		**ŋǀ**	**ŋǃ**			ŋǂ			
	tenuis lateral				**ǁ**						
	voiced lateral				ɡǁ						
	nasal lateral				**ŋǁ**						
Implosive		**ɓ**			ɗ		ᶑ	ʄ	ɠ	ʛ	

Co-articulated	
Labial-alveolar nasal	n͡m*
Labial-velar nasal	ŋ͡m*
Labial-alveolar plosive	t͡p* d͡b*
Labial-velar plosive	k͡p ɡ͡b
Uvular-epiglottal plosive	q͡ʡ*
Labial-palatal approximant	ɥ̊* ɥ
Labial-velar approximant	ʍ **w**
“Swedish sj”	ɧ*
Velarized alveolar lateral approximant	ɫ

IPA coverage: vowels

	Front	Central	Back
Close	i y	ɨ ʉ	ɯ u
Near-close	ɪ ʏ		ʊ
Close-mid	e ø	ɘ ɵ	ɤ o
Mid	e̞ ø̞	ə	ɤ̞ o̞
Open-mid	ɛ œ	ɜ ɞ	ʌ ɔ
Near-open	æ	ɐ	
Open	a ɶ	ä	ɑ ɒ

Setup (so far)

Datasets

Common Voice

- **Japanese** (Japonic)
- **Polish** (Slavic)
- **Maltese** (Semitic)
- **Hungarian** (Ugric)

Total ~20,000 samples

Fine-tuning

Pretrained Model

wav2vec2-large-xlsr-53
Wav2Vec2ForCTC

Additi mall)

Forvo

- **Xhosa** (Bantu)
- **Zulu** (Bantu)
- **Adyghe** (NW Caucasian)
- **Arabic** (Semitic)
- **Icelandic** (Germanic)
- **Burmese** (Tibeto-Burman)

Ran out of time 😢

2000 km
1000 mi
Leaflet | © OpenStreetMap contributors

Results (Example)

Trained on Japanese, Polish, Maltese, Hungarian

Reference (ja):

森永のおいしい牛乳は濃い青色に牛乳瓶をあしらったデザインのパック牛乳である

[moɾinaɡanooiɕiːɡjɯːɲɯːwakoiaoiɾoɲiɡjɯːɲɯːbiɴoaɕiɾatːadezainːopakːɯɡjɯːɲɯːdeaɾɯ]

Prediction:

[moɾinaɡanoːiʃiːɡjɯːɲɯːakoljːaojoɲiɡjɯːɲøːbinoaʃiɾaptavøwainːopakːoɡjɯɲɯːdeaɾu]

Character Error Rate: ~0.231

… Good or bad?

New Metric: Phone Distance (PhD)

[moɾinaɡanooiɕiːɡjɯːɲɯːwakoiaoiɾoɲiɡjɯːɲɯːbiɴoaɕiɾatːadezainːopakːɯɡjɯːɲɯːdeaɾɯ]

[moɾinaɡanoːiʃiːɡjɯːɲɯːakoljːaojoɲiɡjɯːɲøːbinoaʃiɾaptavøwainːopakːoɡjɯɲɯːdeaɾu]

Some IPAs are different but they sound **very similar**

We need a new metric to measure the **phonetic similarity**

Phone Distance (PhD):

Levenshtein Distance with **phonetic features**

e.g., [t] = [-voiced], [+alveolar], [+plosive], …

Averaged → **Feature-based Phone Error Rate** (FPER)

CER: ~0.231, **FPER: 0.122**

Demo

Audio provided by Lusine Vanyan; not for distribution

Karabakh Armenian

Prediction by our model:

* stands for [UNK]

[ɛɾɾku aħbɛɾ an inːum min a fɾɛlunmk ɛnː minːɛl bongi ħiluk ʔaħpɛɾa mjiːʃt ʔɛndon glɛnɛ pinit͡sːnuma ut*ʃ ħt*caɾom ɛnkanad*ʒ t*ʃalum vɛɾ dungiːn ħu isːok odːɾuma min ʔuɾ ɛl jiɾagɛnon ta]

Human transcription (5 min 34 sec):

[eɾk’u ɑχpeɾen inom minə χelũk’ emːinːel tõgi χelũk’ ɑχpeɾɐ miʃt en tõglen binɪt͡snumɐ ut͡ʃʰ t͡ʃɑɾom vɛɾdũŋgin husːə kɑtɾumɐ min oɾel jeɾɑkenum tʰɑː]

(Transcription in Armenian (transliterated): Erku akhper yn njum. Miny khelunk, en miny dongi. Khelunk akhpery en danglen pinycnum a chyrcharum, ver dongin hujsy ktrum a, min or el jer a kenum, ta

Which looks better? CER: ~0.672

FPER: ~0.277

Takeaways

- **New Speech-to-IPA** model
- **Low-resource but good**
- **Faster** language documentation
- **New metric** for IPA generation
- Lots of future work!