Modeling Non-Native Pronunciation

Alexander Metzger | Aruna Srivastava

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

48% of all foreign speakers *struggle* with their accent.

💡 **Goal:** use phonetics to provide foreign speakers granular pronunciation feedback

Issue: most transcriptions are highly ***inaccurate*** *for* ***non-standard speech***

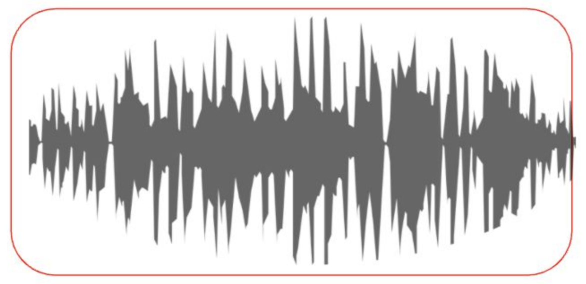
native speech 🇺🇸

Ground truth: **kalɪŋ kaɹdz**

kəlɪŋ kuɹdz

Error: 10%

Non-native speech 🇩🇪



Ground truth: **kolɪ karts**

kolɪŋ kadz

Error: 45%

Question

How can we improve a pretrained model that overfits to representing standard speech to transcribe non-standard speech phonemes, including L2 accented speech and speech impediments?

1) What is the **smallest set of speech sounds** needed to capture both **standard and non-standard** American English?

2) Does following a **curriculum** approach using computer-labeled data **prior** to limited human-labeled data improve performance?

Prior Pretrained Model

XLSR model
500k hours of speech

ASR Intermediary
50k hours of speech

G2P Intermediary
10k hours of speech

G2P Model = BAD

G2P trained checkpoint is ineffective to model non-standard speech well

Gibberish in the pretrained model vocab: 😳

iou4 | ẽʊ̃ | dʑʲ | a.:

Redundancy in the pretrained model vocab :

ɚ = əʴ = əɻ ≈ ɹ

Why???

Data

We curate **~100 hours** of speech data across **9** representative annotated datasets

Dataset Vocab Refinement

Phonetic annotation **varies** across **every dataset and linguistic** annotator.

Datasets use **different phonetic alphabets**

@` = ɚ = AXR

Datasets use **annotators with different training**

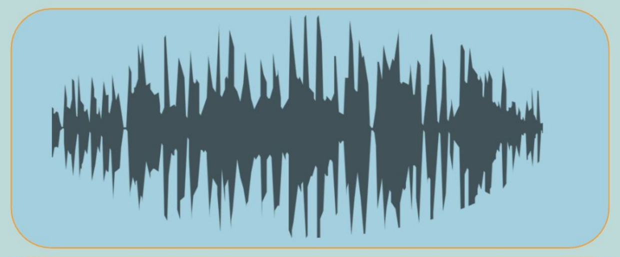
p & pʰ vs p

Datasets use **niche symbols**

ŋ̍ | ɡɣ | β

We must standardize to IPA, collapse redundant phones, and manually remove highly bias/ambiguous samples

G2P Intermediary



Transcription: **Calling Cards**

Grapheme to Phone: **kolɪ karts**

Our Method

Use high-quality **human annotated data** as the **final layer** of learning for phonetic transcription

XLSR model
500k hours of speech

ASR Intermediary
50k hours of speech

G2P Intermediary
10k hours of speech

Human Data
100 hours of speech

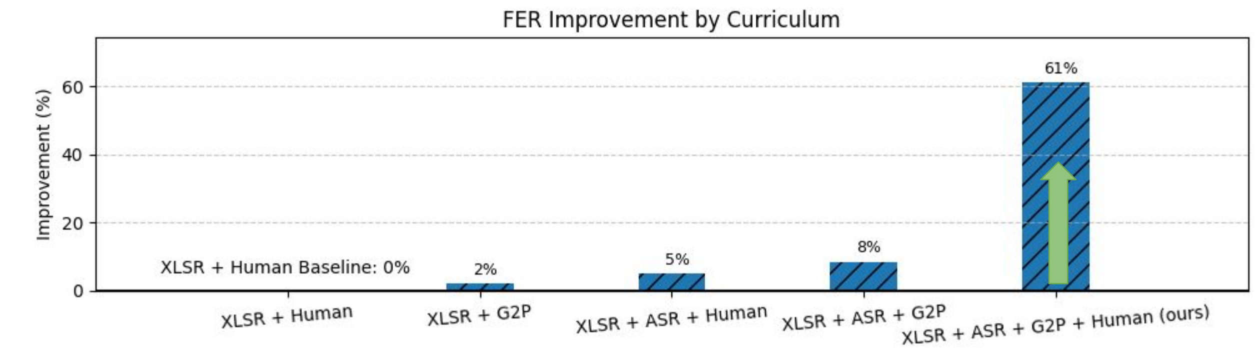
Evaluation

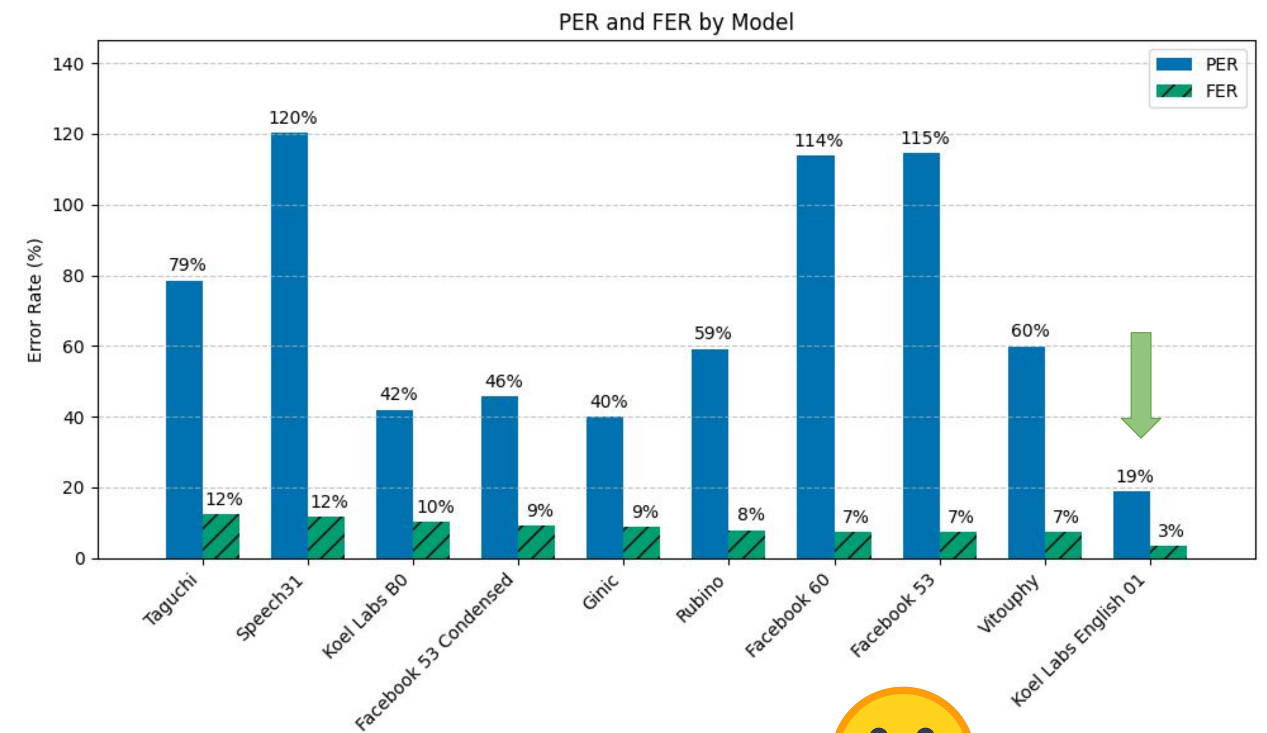
We use phoneme error rate (**PER**) and weighted phoneme error rate (**FER**) for evaluation.

PER considers all character differences equally, FER considers the differences by linguistic distance. For a model that has a similar vocabulary to the test set, it's PER will be superficially high compared to a model that has a slightly different vocab/phoneme notation.

PER		Weighted PER	
P O P		P O P	
B O P	-1	B O P	-.25
I O P	-1	I O P	-.95

Results





Open Sourced! 🤗

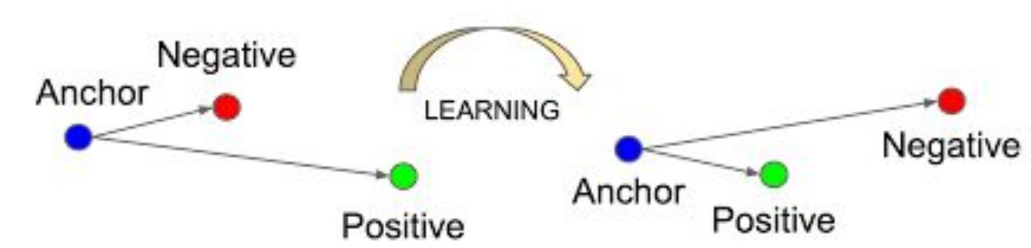
You can test our State-of-The-Art model yourself! https://shorturl.at/f8Y1E

Limitations

- Wav2Vec2 architecture constraints
- Speaker diversity bias
- Unaddressed annotator bias

Future Work

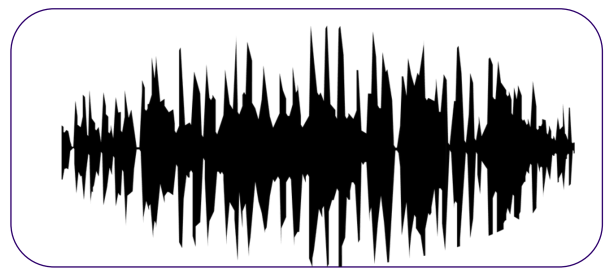
- Handle annotator bias
- Explore IPA representations
- Employ contrastive learning!



Interested in learning more? Sign up for the Beta!

sounds in the pretrained model

native speech 🇺🇸

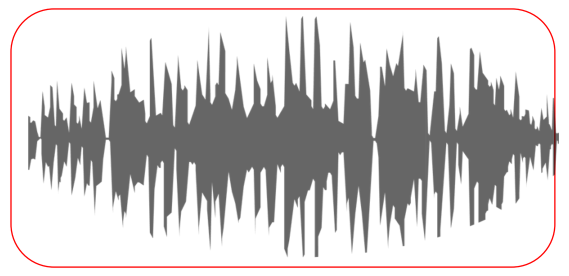


Ground truth: **kalɪŋ kaɹdz**

kəlɪŋ kuɹdz

Error: 10%

Non-native speech 🇩🇪



Ground truth: **kəlɪ karts**

kolɪŋ kadz

Error: 45%

ɚ = əʴ = əɻ ≈ ɹ

Gibberish in the pretrained model vocab: 😵‍💫

iou4 | ẽʊ̃ | dʑʲ | a.ː

Redundancy in the pretrained model vocab :

ɚ = əʴ = əɻ ≈ ɹ

Why???

Datasets use **different phonetic alphabets**

@\` = ɚ = AXR

Datasets use **annotators with different training**

p & pʰ vs p

Datasets use **niche symbols**

ŋ̍ | gɣ | β

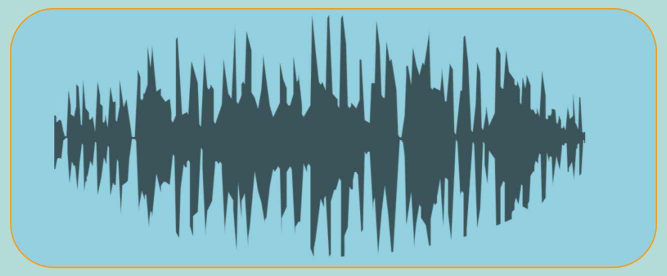
XLSR model
500k hours of speech

ASR Intermediary
50k hours of speech

G2P Intermediary
10k hours of speech

Human Data
100 hours of speech

G2P Intermediary



Transcription: **Calling Cards**

Grapheme to Phone: **kəlɪ karts**